How Training Algorithms Shape Learning: Comparative Recommendations and Insights

Pablo de Vicente
Master in Data Science and AI
Antwerp University
Antwerp, Belgium
Email: pablo.devicenteabad@student.uantwerpen.be

***Abstract*—This paper will delve into different algorithms for training models and how is the outcome affected. Precisely we will be looking into recommender systems for clothes. The daset used consists of 100.000 items of clothes from the well known store hm. We will analyze the algorithms and its limitations, as well as evaluate the results obtained with the use of graphical means.**

1. Introduction

Recommendation systems have become essential to various online platforms, enabling users to seamlessly discover relevant content tailored to their interests. These systems bridge the gap between billions of users and a vast collection of items, ranging from millions to billions of options. While the scale of these systems is immense, the underlying principle remains constant: understanding user preferences and matching them accordingly.

In this paper, we delve into the performance of different training algorithms for recommendation systems, analyzing their outcomes to gain insights into their effectiveness. We employ Recpack, a widely used library for benchmarking recommendation algorithms, and adapt it to suit our specific research goals. Specifically, we evaluate the Popularity, ItemKNN, and KUNN training algorithms, examining their ability to generate personalized recommendations.

To provide a concrete context for our analysis, we utilize a dataset of 100,000 different clothing items from the popular fashion retailer HM. We consider the resulting recommendations from each algorithm, focusing on various aspects of clothing items, such as their category (men, women, sports, etc.), color, price range, fabric, and materials. By analyzing these factors, we aim to assess the level of personalization achieved by each algorithm.

The insights gained from this research will contribute to the development of more sophisticated and effective recommendation systems, enhancing the user experience across various online platforms, particularly in the realm of fashion.

pdv
December 21, 2023

Figure 1. Dataset sample

1.1. Research methodology

To effectively evaluate the performance of the three recommendation algorithms, we employed a rigorous methodology encompassing data preprocessing, algorithm implementation, and metric analysis.

1.1.1. Data Preprocessing. The HM dataset, underwent preprocessing to ensure its consistency and suitability for analysis. This process entailed handling missing values, cleaning data inconsistencies, and extracting relevant features from the raw data. While some of these features were not utilized in the final analysis, others, such as fabric, were successfully extracted from item descriptions.

1.1.2. Algorithm Implementation. We employed the three recommendation algorithms – Popularity, ItemKNN, and KUNN – within our experimental setup. These algorithms were implemented using the Recpack library, a widely used tool for benchmarking recommendation algorithms. The Popularity algorithm chooses an item and recommends it to everyone who has not bought it, if so, it will recommend the second most popular item. The ItemKNN algorithm recommends items similar to those previously purchased or interacted with by a specific user. The KUNN algorithm, an extension of K-Nearest Neighbors, utilizes a combination of user-item interactions and item features to generate recommendations.

1.1.3. Metric Analysis. Traditional metrics like precision, recall, and F-score were employed to evaluate the accuracy and relevance of the generated recommendations. These metrics assess the ability of the algorithms to recommend items that a user would genuinely like. However, we recognized that these metrics do not adequately capture the question of **what** is it that is being recommended.

To address this limitation, we conducted a comprehensive analysis of the distribution of recommendations across various categories, colors, price ranges, fabrics, and materials. These analyses enabled us to gauge the level of personalization achieved by each algorithm. By examining the diversity and relevance of recommended items across these dimensions, we gained insights into the algorithms' ability to tailor recommendations to individual user preferences.

In summary, our research methodology involved rigorous data preprocessing, accurate algorithm implementation, and comprehensive metric analysis. We emphasized the importance of personalization by deviating from traditional metrics and focusing on the distribution of recommendations across various item attributes.

1.2. Dataset

The dataset was obtained from Kaggle submissions and consists of four main components:

- **Articles:** This CSV file contains information about 100,000 items, including their article ID, product name, product group name, color group, department name, and other 20 additional fields that are not essential to this analysis.
- **Customers:** This CSV file provides details about 1.4 million customers, including their customer ID, age, postal code, and information related to their subscription plans.
- **Transactions:** This is the largest CSV file, spanning 40 million records over a two-year period. It captures purchase transactions between customers and articles, with both customer and article IDs identified in previous files.
- **Images:** Zip file containing all images of articles.

1.2.1. Data Limitations. The dataset provides valuable insights into customer behavior and item preferences, but it also has limitations that should be considered. Firstly, the dataset covers only two years of transactions, which may not fully reflect the dynamic nature of customer preferences and item popularity over time. Due to computational constraints, we opted to utilize a subset of the transactions, focusing on the last 11 weeks of our sample. This decision allowed us to effectively analyze the dataset while maintaining manageable computational demands. Our testing and validation dataset was structured as follows:

- final date: 2020-09-22
- t date : 2020-09-15
- delta in: 10 weeks (10 * 604800s)
- delta out: 1 week (1 * 604800s)

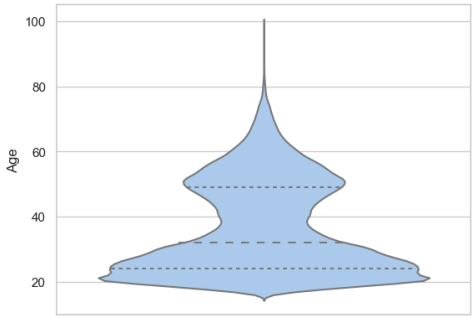


Figure 2. Age distribution

The period spanning 10 weeks prior to the t date was designated for testing purposes, ensuring that the recommendations generated by the algorithms could be evaluated against actual user behavior. The period spanning 1 week after the t date was set aside for validation, allowing us to assess the algorithms' ability to generalize to new user interactions. This approach enabled us to balance the need for a comprehensive dataset with the computational limitations while maintaining the rigor of our evaluation.

1.3. Preliminary Analysis of dataset

In order to grasp our initial conditions for the dataset we will do a basic exploratory analysis of both users and items.

1.3.1. Users. The dataset offers limited insights into user demographics beyond age. However, the age distribution, depicted in Figure 2, reveals two distinct groups of active buyers. The first group comprises individuals around the age of 20, while the second group consists of those around 45 years old.

1.3.2. Transactions. The transaction data offers insights into customer purchase behavior, but it doesn't reveal any particularly noteworthy patterns or trends that would significantly impact our recommendation algorithms. As a result, we won't focus on this aspect in our analysis.

1.3.3. Articles. The article data provides a more detailed view, allowing us to examine the distribution of clothing categories. As revealed in Figure 3, ladieswear comprises a substantial portion of the dataset. This trend can be attributed to the fact that women tend to be more active shoppers, sadly, there is no information on customers as to their gender.

Another interesting aspect is the distribution of colors, as illustrated in Figure 4. To enhance the visual appeal and simplify the interpretation of the graph, we have employed a color scheme that groups closely related hues together. This approach helps to minimize visual distractions and improve the overall readability of the graph. Unsurprisingly, black emerges as the most prevalent color.

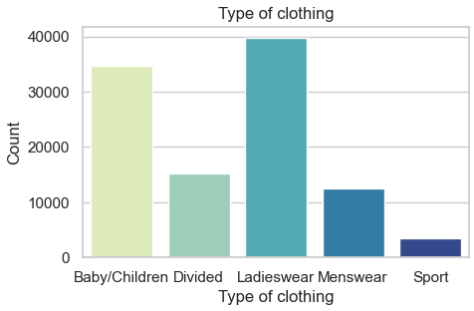


Figure 3. Clothes distribution

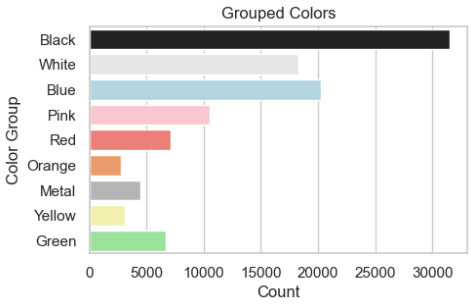


Figure 4. Color distribution

Even though there are more aspects of the dataset to analyze, we've determined not to go into detail, as they will come up later on while doing the studies of results.

2. Results

In this section of the paper we will discuss the results obtained by the different training algorithms

2.1. ItemKNN

Also called k-nearest neighbors, ITEMKNN is an association rule learning algorithm that relies on the k-nearest neighbors (kNN) concept. It identifies association rules between items based on the similarity of co-purchased items

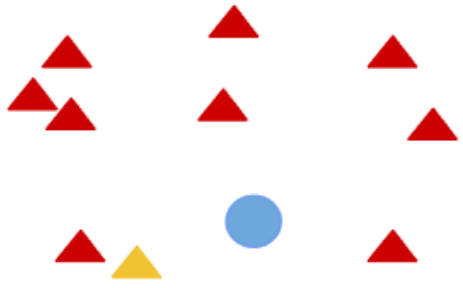


Figure 5. Visual representation of ITEMKNN. The closest triangle (representing our items) to our query (represented as the blue dot) is the orange one, thus, it'll be the recommended item

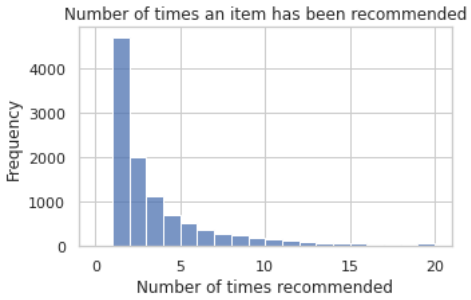


Figure 6. Number of times each article has been recommended

within customer baskets. Figure 5 provides a representation of the model. By measuring the distances from our query to each of the items we can choose the one closest, making it the most suitable candidate for our recommendation. It is important to note that similar items will be grouped closer together.

For our testing we evaluated

$$k@\{100, 200, 500\}$$

and for similarity we both used cosine and conditional distances

$$similarity = \{\texttt{cosine}, \texttt{conditional\_probability}\}$$

2.1.1. Analysis ITEMKNN. Our validation data consists of the last week of available data, spanning from September 15th to September 22nd, 2020. During this period, there were approximately 75,000 transactions made. After removing users who have made less than two transactions, we are left with 45,000 customers to generate recommendations for

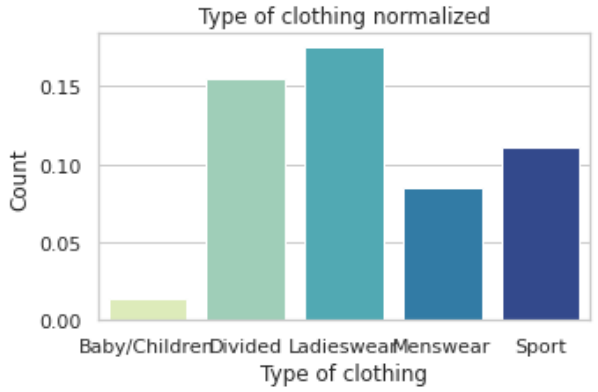
Figure 6 illustrates the distribution of items based on the number of times they have been recommended. The most frequent recommendation count is 4, which makes sense as we have generated 11,000 recommendations for 45,000 customers.

When looking at the distribution of recommended clothing types (Figure 7(a)), Ladieswear overwhelmingly dominates the recommendations. This is likely due to the dataset's heavy skew towards Ladieswear as we already observed in 3. To account for this bias, Figure 7(b) normalizes the recommendations, which balances out the distribution somewhat, with the exception of Babies clothes, which are not being recommended at all. We will theorize on why this is in the conclusions section.

The model's recommendations exhibit a balanced distribution across color groups, with no apparent preference for any particular color scheme. Notably, there's a noticeable avoidance of blue colors, favoring a more prevalent black/white palette. Additionally, while not directly presented, The distribution of tones leans towards darker shades rather than lighter ones.

The next part of the analysis was an exploratory endeavor. Fashion blogs often suggest that specific colors are associated with different seasons, with brown-ish tones

(a) Image 1: number of articles recommended



(b) Image 2: number of articles recommended normalized

Figure 7. Comparison in recommended articles.

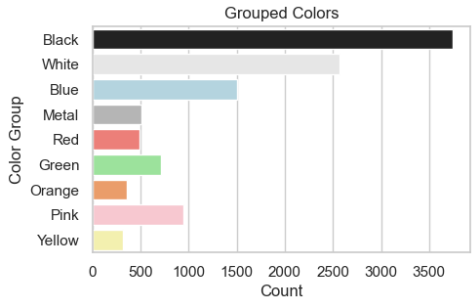


Figure 8. Colors recommendation grouped

prevalent in autumn, vibrant colors like red, yellow, and green dominating summer, and darker hues taking center stage in winter. As our data was collected during the third week of September, we were curious to see if this color theory aligns with our findings. As depicted in Figure 9, winter colors emerged as the most recommended category (likely due to the inclusion of black and white). Following closely behind were summer colors, which makes sense considering that our training data primarily comprises items from the summer months. Autumn colors trailed behind, showcasing the influence of season on color preferences.

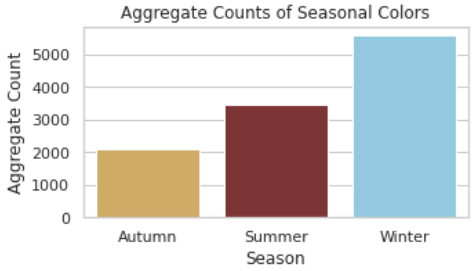
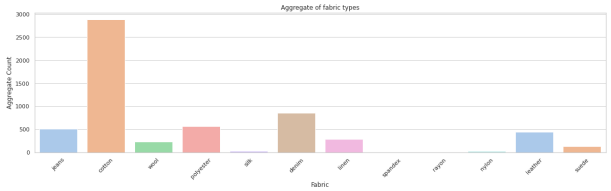


Figure 9. Colors grouped by seasonality

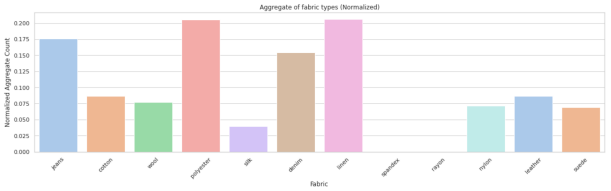
The color groupings follow common color trends that are often associated with different seasons, as an example, Winter is associated with dark and neutral colors like black, white, navy, charcoal, and burgundy. Autumn on the other hand likes earthy tones like brown, tan, olive, and mustard. Lastly summer is closely resembled by vibrant colors like red, orange, yellow, green, and pink. Bellow there is a table specifying which color has been grouped to each season.

TABLE 1. COLOR GROUPINGS

Season	
Winter	Turquoise, Lilac Purple, Mole, Grey, Blue, Black, White
Autumn	Red, Metal, Orange, Brown, Beige
Summer	Bluish Green, Yellowish Green, Green, Khaki green, White, Pink, Yellow



(a) Image 1: Distribution of fabric types



(b) Image 2: Distribution of fabric types normalized

Figure 10. Comparison in fabric recommendations

Our next analysis focused on fabric, a feature that we manually extracted from the product descriptions. Since this information is not directly available from the original dataset, the accuracy of fabric classification is limited. Despite this limitation, we observed a strong prevalence of cotton as the most common fabric, which aligns with its widespread use in clothing. This is reflected in Figure 10(a), which shows the distribution of recommended fabrics.

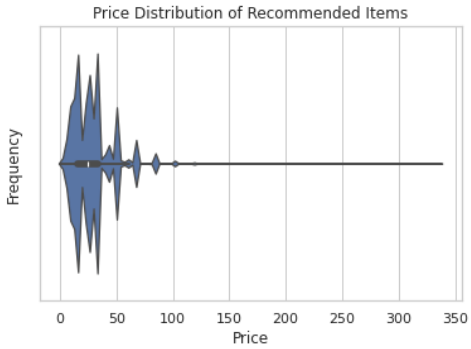


Figure 11. Price distribution of knn recommendations

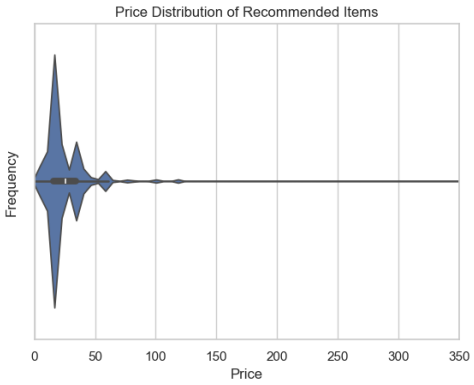


Figure 12. Price distribution of all articles

Intriguingly, when we normalize the fabric recommendations (Figure 7(b)), we discover a surge in four specific fabrics: linen, polyester, jeans, and denim. This unexpected pattern suggests that these fabrics may hold particular appeal to the target audience or that they are particularly well-suited to the current season or trends.

Finally, we turned our attention to price, a crucial aspect of clothing recommendations. However, the price values in the dataset are expressed in small decimals, making them difficult to interpret directly. To gain a better understanding, we multiplied all price values by a constant C (in our case, 1000) to convert them into more meaningful units. This transformation shifts the price distribution to a range more familiar to consumers, ranging from 10 to 60 euros predominately.

Figure 11 reveals several noteworthy patterns in the price recommendations. The most prominent peaks occur at prices of 10, 20, 30, 40, 50, and 60 euros. Interestingly, the peak at 40 euros is relatively smaller, suggesting that recommendations at either 30 or 50 euros may be influencing this category. This observation could indicate price thresholds or preferences among the target audience.

Finally, we compared the price distribution of recommended items to the overall price distribution of clothing items in the dataset (Figure **??**). This comparison reveals a significant skew towards the more expensive side for

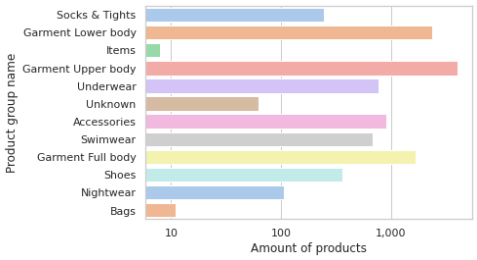


Figure 13. Division of recommendations by clothes type

recommended items. This suggests that the recommendation system is tending to favor higher-priced items, potentially reflecting the target audience's preference for more premium or stylish clothing pieces.

2.1.2. Conclusion Recommendations ItemKNN. After thoroughly examining the data, we'd like to highlight some observations that haven't received the attention they deserve. These insights aim to shed light on the inner workings of the recommendation system, which remains an intriguing puzzle. While we've gained valuable insights, several questions remain unanswered, prompting further exploration and analysis.

- The model's underrepresentation of babies/children's clothing despite its significant presence in the dataset raises an interesting question. One possible explanation is that recommendations for children's clothing are more likely to be wrong, leading to a higher penalty for the model. To avoid these penalties, the model may be erring on the side of caution by recommending more general options, such as Ladieswear and Divided, which cater to a wider audience.
- The prevalence of linen, jeans, and polyester in recommendations, even though cotton is the most common fabric, stands out as an intriguing pattern. Initially, we hypothesized that this might be due to the model's preference for lower garments, as these fabrics are commonly used for said items. However, this hypothesis was disproven by Figure 13, which shows a distribution of lower garments in line with expectations. This leaves us with an unresolved mystery regarding the increased prevalence of these specific fabrics.

2.2. KUNN

KUNN (Unified Nearest Neighbour) is a hybrid similarity recommendation algorithm that utilizes both user-based and item-based collaborative filtering techniques. It aims to leverage the strengths of both approaches to generate more accurate and personalized recommendations.

User-based collaborative filtering considers the similarities between users to recommend items that similar users

have liked or purchased. This approach works well when there are strong patterns in user preferences.

Item-based collaborative filtering, on the other hand, considers the similarities between items to recommend items that are similar to those the user has previously interacted with. This approach is effective when there are strong patterns in item characteristics.

KUNN combines these two approaches by creating a unified neighborhood that includes both similar users and similar items. This allows the algorithm to take into account the preferences of similar users and the characteristics of similar items when making recommendations.

The algorithm works by first calculating the similarity between all users and items based on their historical interactions. Then, it identifies a set of k nearest neighbors for each user and item. These neighbors represent the users and items that are most similar to the target user or item.

To generate recommendations for a particular user, KUNN calculates the weighted average of the ratings given to each item by the user's k nearest neighbors. The weights are based on the similarity between the user and their neighbors. This ensures that items that are liked by similar users are given more weight in the recommendation.

Similarly, to generate recommendations for a particular item, KUNN calculates the weighted average of the preferences of the item's k nearest neighbors. The weights are based on the similarity between the item and its neighbors. This ensures that items that are similar to other popular items are given more weight in the recommendation.

KUNN has been shown to outperform both user-based and item-based collaborative filtering approaches in several studies. This is because it can effectively combine the strengths of both methods to provide more accurate and personalized recommendations.

Here are some of the key advantages of KUNN:

- Improved accuracy: KUNN can achieve higher recommendation accuracy than both user-based and item-based collaborative filtering approaches.
- Better personalization: KUNN can provide more personalized recommendations by considering both user preferences and item characteristics.
- Scalability: KUNN is scalable to large datasets and can handle a large number of users and items.

Overall, KUNN is a powerful recommendation algorithm that can be used to generate accurate and personalized recommendations for a wide variety of applications, including our recommender system. However, it is important to note that KUNN requires a high amount of computational power, as it needs to calculate the similarity between all users and items in the dataset. This can be a significant challenge for large datasets, and it may require the use of specialized hardware or cloud computing resources. Such is our case that we were not able to run the model for a significant enough portion of data, if the reader recalls from section 1.2.1, we where initially using 11 weeks of data to train/test our model, to run KUNN with this sample size it was required around 800GiB of free memory (which of course, the testing PC did not have). To this extent, we started reducing the amount of test/train data and at some point we where able to run and train the model. Although we could train it, it was with such a small sample size that we dont feel our model is general enough, rather, it depended on what specific week of data it was inputing in order to generate recommendations. For that matter, even though we could, we did not feel like incorporating the plots and results for such models, as it dosent closelly resemble the vision of a general algorithim and it is mostly the dependent on each iteration.

Figure 14. Top three recommended items from left to right

2.3. Popularity

Based on our understanding of Radek's notebook, we initially assumed that the popularity metric would group users into specific "stereotypes" and recommend the most suitable item for each group. However, Recpack's popularity algorithm deviates from this approach. Instead, it generates a list of items ranked by popularity, identifying the item that is most likely to be universally appealing. It then recommends this top-rated item to every user. If a user has already purchased this item, the algorithm proceeds to recommend the second most popular item from the list. This approach explains why Recpack has recommended the same pair of pants (shown in Figure 14) to over 44,500 users. For the approximately 480 users who had already purchased these pants, the algorithm recommended a different pair of pants instead.

In line with the insights from previous sections, the recommended items consistently fall into specific categories: they are predominantly black, belong to the Ladieswear or Divided sections, and are primarily made of jeans fabric.

3. Conclusion

Although the paper has been able to provide some insights into recommendation algorithms, it is not as much as we expected, the fact that we werent able to train a KUNN model with enought data saddens us. Either way we believe that at least we have gained some knowledge on the process of training models, and some proficiency with analyzing and finding relations data, labor that was way harder than expected.

In conclusion, this paper has offered a glimpse into the world of recommendation algorithms and their application

in fashion e-commerce. While we were unable to fully explore the potential of a KUNN model due to computational limitations, our exploration of popularity metrics and item characteristics has yielded valuable insights.

Our analysis revealed that Recpack's popularity algorithm prioritizes items that are widely liked and universally appealing, often recommending the same pair of pants to over 44,500 users. This approach, while effective in reaching a wider audience, may overlook the unique preferences of individual users.

On the other hand, our analysis of item characteristics highlighted the prevalence of black, ladieswear, and divided items in recommendations. This aligns with the traditional color and style trends associated with these categories. However, it raises the question of whether the algorithm is effectively catering to the diverse preferences of its users.

Despite these limitations, our work has provided valuable insights into the workings of recommendation algorithms and the challenges associated with analyzing large datasets. We have gained hands-on experience in training models and extracting meaningful patterns from data, skills that are essential for working with recommendation systems.

Overall, while this study hasn't been as conclusive as the author would have liked, it has opened up new posibilities for investigation and yielded the necessary knowledge to delve further into related fields.

4. Further improvements

Due to time constrains, our work did not aproach every aspect of training a model and there is still improvemts to be made, in this section we will list a few.

- The dataset primarily focuses on purchase behavior, which may overlook other aspects of customer engagement, such as browsing history, items added to the cart but not purchased, or the time spent interacting with specific items. These limitations can hinder the ability to generate accurate negative samples, which are crucial for improving recommendation performance.
- Users without buying history should be presented with basic options to choose from, this problem is usually addressed as "cold start users". In our implementation we got rid of any user/item that didn't have at least 2 interactions.
- To anyone with sufficient computational power, it would be interesting to observe the results of a KUNN model with enough training data to make it reliable

References

[1] camillestyles, title = CamilleStyles — Color Analysis, url = https://camillestyles.com/beauty/color-analysis/, lastchecked = 2023-12-22,

[2] camillestyles2, title = CamilleStyles — Autumn Color Analysis, url = https://camillestyles.com/beauty/fashion/autumn-color-analysis/, lastchecked = 2023-12-22,

[3] recpack, title = RecPack — Froomle AI, url = https://recpack.froomle.ai/index.html, lastchecked = 2023-12-22,

[4] wu2017starspace, title=StarSpace: Embed All The Things!, author=Ledell Wu and Adam Fisch and Sumit Chopra and Keith Adams and Antoine Bordes and Jason Weston, year=2017, eprint=1709.03856, archivePrefix=arXiv, primaryClass=cs.CL

[5] Ni, Jingwen, title = UChicago Knowledge, url = https://knowledge.uchicago.edu, lastchecked = 2023-12-22,

[6] yi2019sampling, title=Sampling-bias-corrected neural modeling for large corpus item recommendations, author=Yi, Xinyang and Yang, Ji and Hong, Lichan and Cheng, Derek Zhiyuan and Heldt, Lukasz and Kumthekar, Aditee and Zhao, Zhe and Wei, Li and Chi, Ed, booktitle=Proceedings of the ACM Conference on Recommender Systems, year=2019, url=https://dl.acm.org/doi/10.1145/3298689.3346996