PROJECT 02807 COMPUTATIONAL TOOLS FOR DATA SCIENCE - GROUP 40

Personalized News Recommendation System

Justine Fastner, Stefanie Lanz, Niclas Pokel, Patrick Schmitt
s242066, s242107, s242105, s241893

GitHub Repository

Abstract

This project aims to recommend news articles customized to users' interests to enable personalized reading experiences and trend discovery. Three methods were investigated: Frequent Items algorithms, including A-Priori and PCY, to identify popular article combinations; content based recommendation systems using various clustering techniques and BERT; reinforcement learning to dynamically adapt to changing user preferences. The project is based on the Microsoft News Dataset (MIND).

1 Introduction

The increasing use of online platforms has made people more selective in choosing news content. Personalized news that match the individual's interests is more likely to be accepted than content sourced from a general platform. Furthermore, it becomes more important to keep up with the latest trends and to be aware of what others are interested in [1]. Personalized recommendation systems are essential for navigating large amounts of news, filtering content, and highlighting relevant topics.
Our solution evaluates three approaches for recommending news articles: identifying popular article combinations, clustering feature groups within articles, and leveraging reinforcement learning for model training. The methods we applied were aligned with the course curriculum and extended beyond it, incorporating advanced techniques. We compared and discussed the results in detail in the latter part of the report. The research questions we addressed are the following:

1. How can frequent item pairs from user histories be used for recommendations with A-Priori and PCY?
2. How can classical clustering methods (Hierarchical Clustering, k-means, DBSCAN) classify article features for recommendations, using both TF-IDF and BERT?
3. Can Deep Q-Learning improve recommendations?

2 Data

2.1 Dataset

We used the Microsoft News Dataset (MIND), which records user interactions with approximately 160,000 English news articles over six weeks. It includes anonymized data from one million users who clicked on at least five articles a week. A subset of 50,000 users was utilized for training, validation, and testing. The dataset contains 15 million impression logs and detailed textual content such as title, abstract, category and subcategory [2]. The two main files, behaviors.tsv and news.tsv, provide user click histories, impression logs, and article metadata [3].

2.2 Data Preprocessing

Our preprocessing involved data cleaning and preparation for clustering. Missing values in categories, subcategories, titles, and abstracts were either replaced with placeholders or removed. We eliminated duplicates in news items and impression logs based on title, abstract, and impression IDs. The textual data was tokenized, with special characters, numbers, HTML tags, and stopwords removed, and the text converted to lowercase. We used stopwords inspired by the NLTK library [4], though the implementation was done independently. Lemmatization was then applied to reduce words to their base form. Figures 1 and 2 show the word distributions before and after preprocessing. To recommend frequent items, we analyzed complete user baskets by grouping user IDs and consolidating their entire reading history into one per user.
For data preparation in clustering, we converted textual features such as titles and abstracts into TF-IDF vectors and combined them with one-hot encoded categorical features for categories and subcategories. This unified feature set was reduced to 100 dimensions using Principal Component Analysis (PCA), which improved efficiency while preserving essential information [5, 6]. We chose PCA over t-SNE due to its ability to preserve the global data structure, which is crucial for clustering tasks [7]. This process was particularly important for the clustering task.

3 Methods

3.1 Content based Recommendation with Clustering

For the content based recommendations, the features were optimized to improve clustering and recommendation performance. With the process from the data preprocessing, a feature matrix of (50669, 100) is obtained with TF-IDF, one-hot and PCA. We applied the clustering methods to this feature matrix. The clustering methods we tested included k-means, DBSCAN and hierarchical clustering. For k-means, the elbow method identified k = 20 as optimal, while the silhouette analysis suggested k = 4. The elbow method is based on the position of the sum of squared errors [8], while the silhouette analysis evaluates the visualization quality of each cluster [9]. We used k-means for recommendations by analyzing each user's history to identify the most frequent cluster of articles they had read. The top ten articles from this cluster, excluding those already read, were recommended to match users' interests.

3.2 Content based Recommendation with BERT

BERT and TF-IDF are both vectorization techniques but differ significantly. BERT creates dense, context-aware vectors by capturing relationships within text, while TF-IDF generates sparse vectors based on term frequency [10]. BERT excels in contextual encoding, making cosine similarity more effective than clustering. Unlike TF-IDF, which can misinterpret context (e.g., linking "2024 US election" articles to older elections due to shared terms), BERT captures nuances such as specific candidates, improving recommendation precision. TF-IDF's hyperparameter tuning may exclude relevant information. Our hybrid recommendation system combines BERT-based content analysis with user interaction history. It computes content similarity by averaging item embeddings that a user has interacted with and calculating cosine similarity with all item embeddings using the DistilBERT model. Collaborative filtering scores are derived by averaging user vectors from interacted items and calculating cosine similarity with item-user vectors. We combine these scores equally for final recommendations, enhancing explainability and enabling further optimization. To manage BERT's memory demands, we store only the top K most similar items for each item, reducing matrix size and computational overhead while balancing contextual understanding and scalability.

3.3 Frequent Items based Recommendation

To find frequent item pairs we implemented the A-Priori and PCY algorithms, using user histories as "baskets" to analyze article readings and combinations. A-Priori offers a comprehensive view of frequent items, while PCY is more efficient but may introduce data loss due to hashing. We selected to base the recommender system on the PCY for its scalability with large datasets. If the input article is part of a frequent item pairs, we recommend those articles. If the article is frequent but not part of enough frequent item pairs, recommendations are based on the candidate pairs. For articles that are not frequent, we recommend the most frequent articles from the same category. This approach leverages PCY's efficiency while ensuring relevant suggestions, however the relevance for non-frequent articles is hard to quantify.

3.4 Deep Q Reinforcement Learning (DQR)

Deep Q-Learning, a reinforcement learning method, trains an agent to optimize long-term user engagement by estimating Q-values, which predict the expected rewards for recommending specific news articles. The model improves recommendations by maximizing click-through rates (CTR) and adapting to user preferences in a simulated environment based on interaction histories and metadata.
Unlike BERT, which focuses on content understanding and static user preferences, DQR adapts dynamically, making it effective for sequential decision-making and online learning. However, it may struggle with cold-start problems and requires more interaction data to converge. While the dataset was large, sparse user interactions could affect recommendation relevance. Another drawback is its low interpretability, as it generates abstract representations of statistics, unlike simpler methods that provide easily understood metrics.

4 Results

4.1 Content based Recommendation with Clustering

The clustering methods show varying strengths and weaknesses, with k-means and t-SNE being the most effective for news article recommendations. Hierarchical clustering offers high accuracy and detailed insights but is impractical due to its long runtime and complexity. Figures 5 and 6 show hierarchical clustering for n_clusters = 4 and 20. K-means, visualized using PCA, reveals clusters with some overlap, and three-dimensional representations highlight greater complexity. Figures 7 and 8 show k = 4 and k = 20 in two dimensions, while figures 9 and 10 depict them in 3D. t-SNE visualizations reveal peripheral clusters and central overlap, as shown in figures 11 and 12. DBSCAN, despite testing various parameters, was less informative, with small eps values producing many outliers and large values grouping most points into one cluster. Thus, k-means and t-SNE provide more meaningful clustering results. The k-means-based recommendation system effectively identifies the top 10 articles from the main cluster, excluding previously read articles, as shown in figure 13.

4.2 Content based Recommendation with BERT

As shown in figure 14 and 15, the model relies heavily on semantic similarity, reflected in much higher content scores compared to collaborative scores. This suggests the presence of many closely related articles (can be confirmed by reviewing entries and recommendations) but only weakly correlated interaction behaviors. For final optimization, the weighting between content and user interactions should be reevaluated. In the current configuration, user interactions provide little relevant information. Thus, for this task, a pure content based recommendation using BERT may suffice without additional interaction data. Overall, the content and context embeddings perform well, producing semantically relevant results, while sparse user interactions in the smaller dataset fail to add meaningful insights.

4.3 Frequent Items based Recommendation

For the frequent item pairs, two graphs illustrate the number of frequent articles, pairs, and, for A-Priori, frequent triplets. Figures 16 and 17 show the results for different thresholds, demonstrating expected behavior. The thresholds indicate the minimum number of readers required for an article or pair to be considered frequent. In figure 18, it is evident that both algorithms identify the same top item pairs. For the recommendations, as shown in Figure 19, 10 articles are recommended, with clear reasoning for each suggestion. This approach remains highly efficient, even when applied to the much larger dataset.

4.4 Deep Q Reinforcement Learning (DQR)

The DQR-Agent was trained to a low MSE by receiving rewards for correctly recommending articles, as shown in figure 20. However, the final test accuracy for recommending the top 5 articles remained 0, likely due to the sparse dataset. A larger dataset with more users and interactions would likely improve performance. Diagnosing issues with such abstract methods is challenging. For example, the DBSCAN clustering approach

showed seemingly random clusters, as seen in figure 21. Figure 22 shows strong alignment between user and article embeddings, which could suggest poor generalization or random structure in low dimensions. A more complex agent resulted in the embedding distribution in figure 23.

In conclusion, while DQR-Learning is theoretically efficient, it struggled to separate distinct distributions, possibly due to the smaller dataset, reduced information density, or insufficient agent complexity and optimization.

5 Discussion

The project highlights the feasibility of combining frequent item analysis, content based clustering and reinforcement learning for personalized news recommendation systems. Only frequent items methods scaled well with the large dataset, as content based clustering and reinforcement learning required more resources and were impractical for large-scale deployment. Content based clustering provided valuable insights but required extensive preprocessing, such as stopword removal and PCA, which limited scalability. BERT improved the understanding of the data but presented a challenge for the analysis of our large dataset. PCY showed efficiency in identifying popular article combinations in large datasets, but was lacking in recommendations for niche topics. DQR showed potential for dynamic user customization, but relied on dense interaction data, which limited its immediate applicability. Hybrid systems, despite their promising concept, had difficulties with metric matching and optimization.

6 Conclusion

This project outlines both the strengths and limitations of various recommendation techniques. Frequent items based recommendations proved to be efficient and scalable, making them well-suited for large datasets. However, they were less effective in addressing niche interests. Clustering methods and BERT significantly enhanced personalization, but their computational demands made them less practical for larger datasets. Reinforcement learning showed potential for improving long-term user engagement, though it requires further development for broader applicability. Future work should focus on enhancing hybrid models, addressing scalability challenges, and optimizing these techniques to better align with diverse user needs and dataset characteristics. Overall, these approaches provide a solid foundation for developing effective, scalable recommendation systems, with room for further improvements.

References

[1] Paul Sagan and Tom Leighton, "The internet & the future of news," *Daedalus*, vol. 139, no. 2, pp. 119–125, 04 2010.

[2] MSNews, "Mind: Microsoft news dataset," `https://msnews.github.io/`, Accessed: 2024-11-12.

[3] MSNews, "Introduction to the mind and mind-small datasets," `https://github.com/msnews/msnews.github.io/blob/master/assets/doc/introduction.md`, Accessed: 2024-11-12.

[4] PythonSpot, "Nltk stop words," `https://pythonspot.com/nltk-stop-words/`, 2024, Accessed: 2024-11-20.

[5] Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya, "Document clustering: Tf-idf approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, 2016, pp. 61–66.

[6] Takio Kurita, "Principal component analysis (pca)," *Computer vision: a reference guide*, pp. 1–4, 2019.

[7] Jyoti Pareek and Joel Jacob, "Data compression and visualization using pca and t-sne," in *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019*. Springer, 2021, pp. 327–337.

[8] Muhammad Ali Syakur, B Khusnul Khotimah, EMS Rochman, and Budi Dwi Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*. IOP Publishing, 2018, vol. 336, p. 012017.

[9] SM Aqil Burney and Humera Tariq, "K-means cluster analysis for image segmentation," *International Journal of Computer Applications*, vol. 96, no. 4, 2014.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

Appendix

List of Figures

Figure 1: Word Clouds Before Data Cleaning

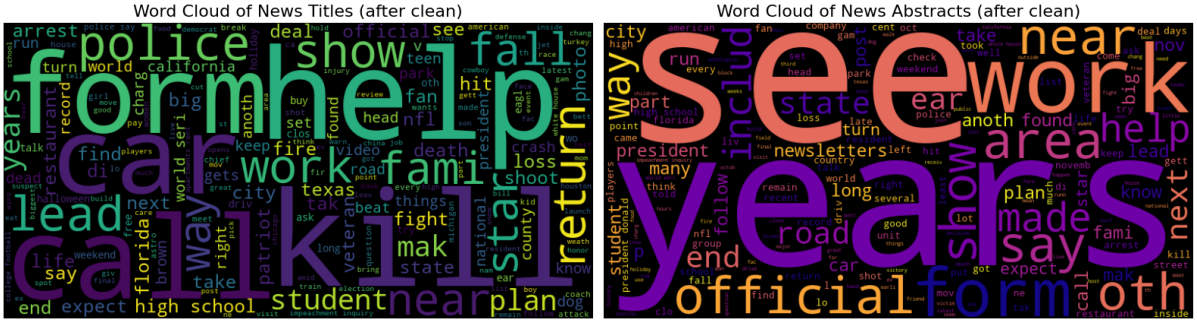

Figure 2: Word Clouds After Data Cleaning

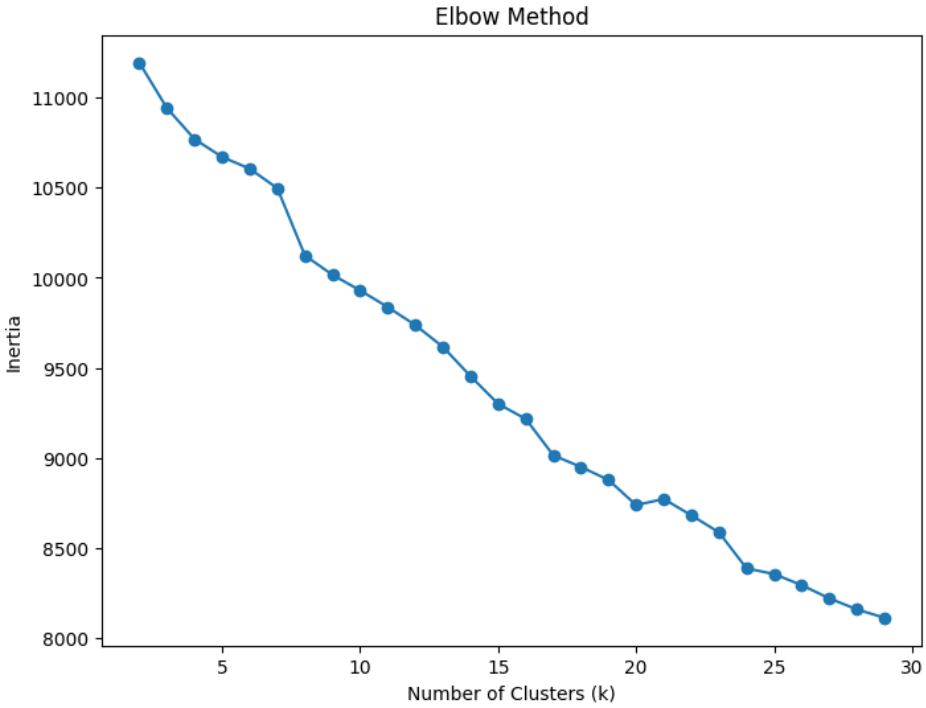

Figure 3: Elbow Method

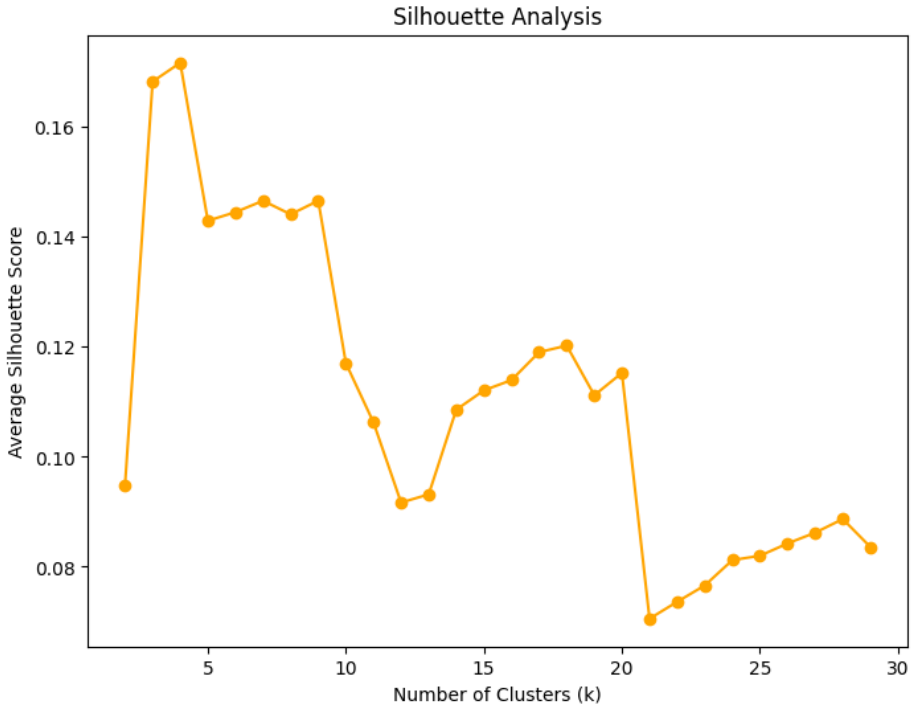

Figure 4: Silhouette Analysis

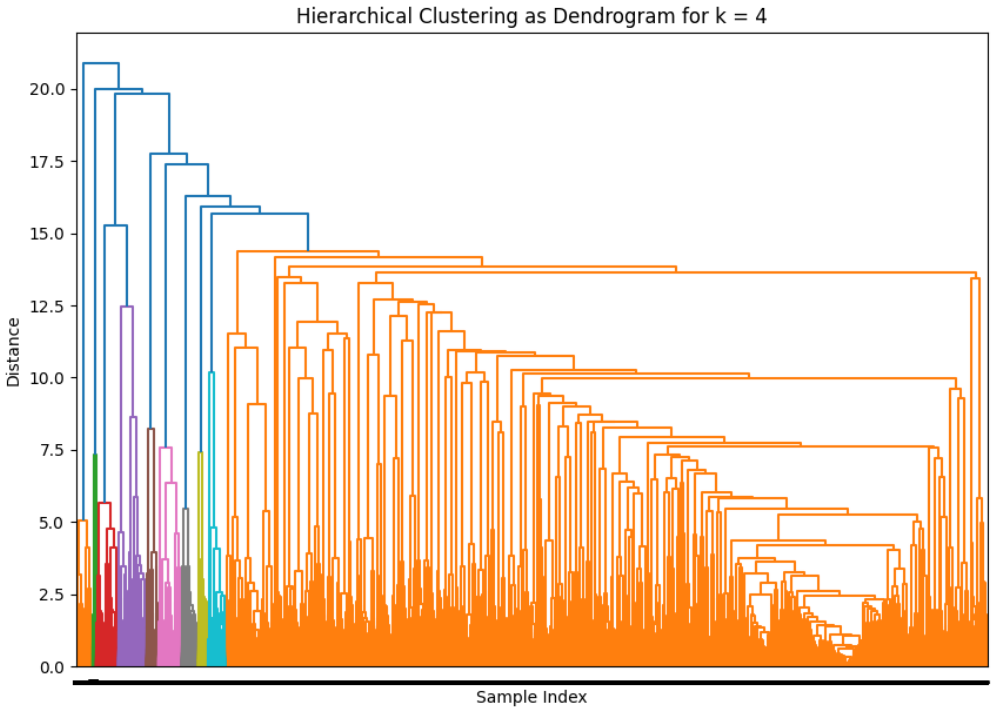

Figure 5: Hierarchical Clustering with k = 4 and n_clusters = 4

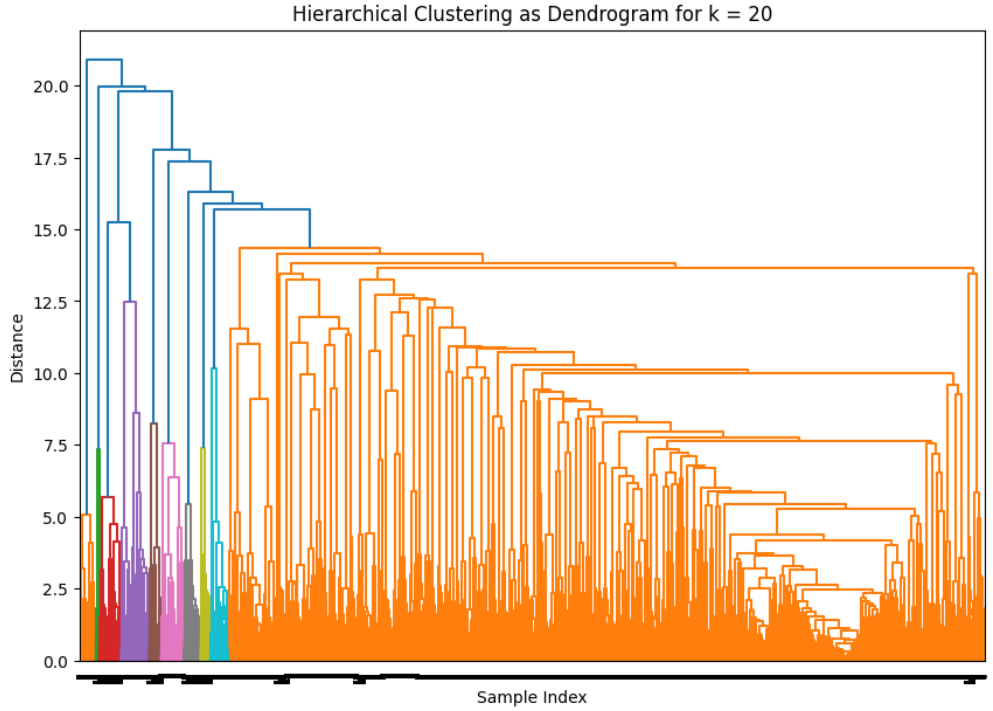

Figure 6: Hierarchical Clustering with k = 20 and n_clusters = 20

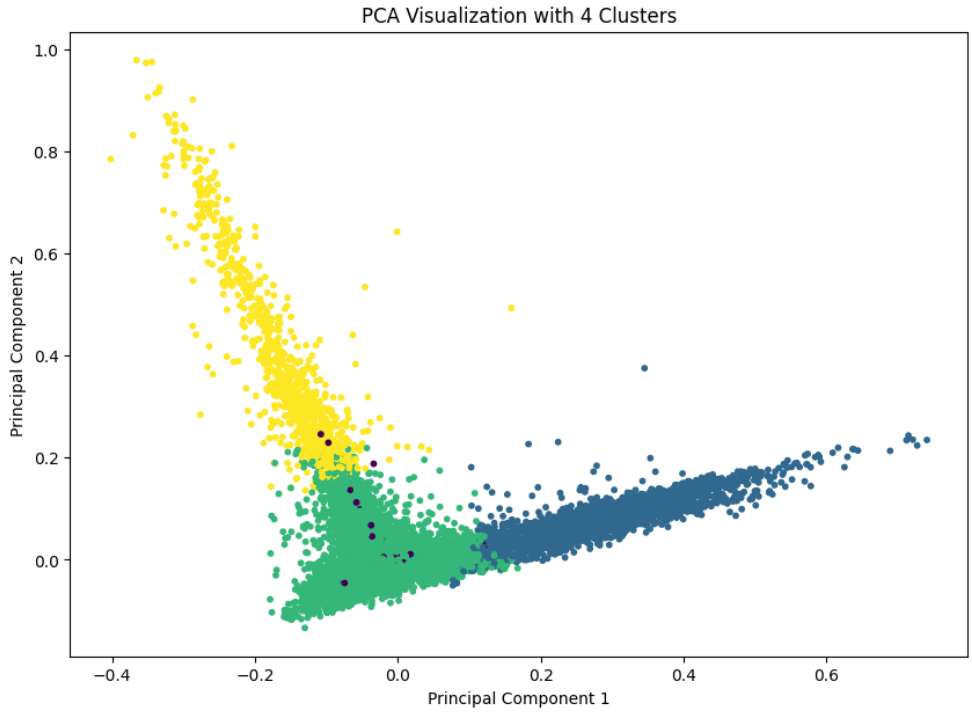

Figure 7: K-means Clustering with k = 4 and PCA Visualization

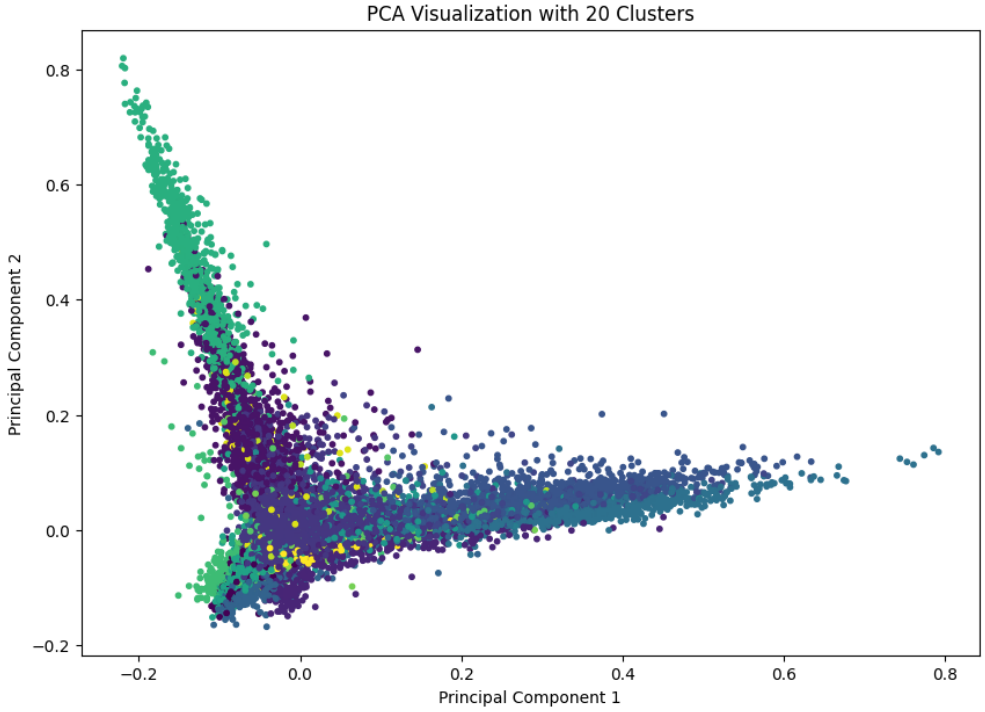

Figure 8: K-means Clustering with k = 20 and PCA Visualization

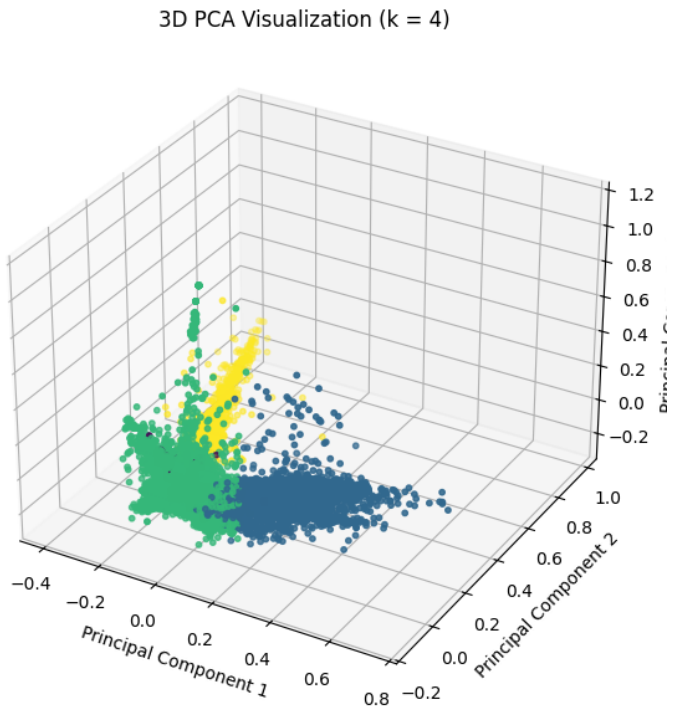

Figure 9: K-means Clustering with k = 4 and 3D PCA Visualization

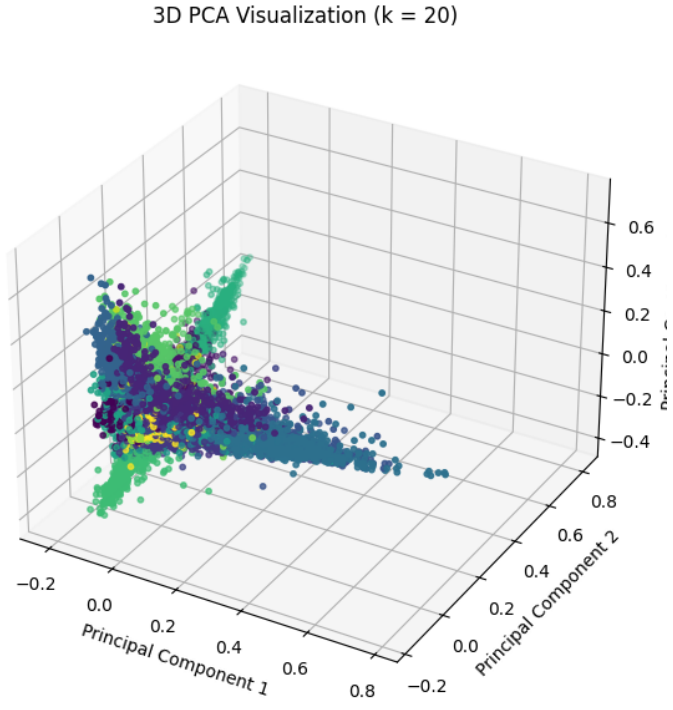

Figure 10: K-means Clustering with k = 20 and 3D PCA Visualization

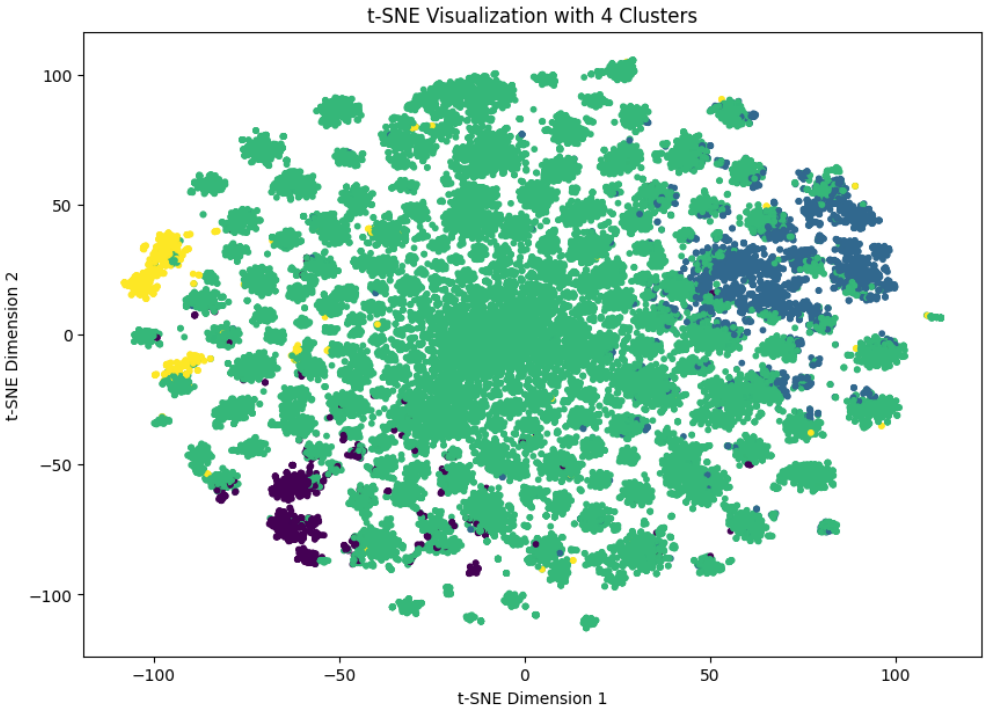

Figure 11: K-means Clustering with k = 20 and t-SNE Visualization

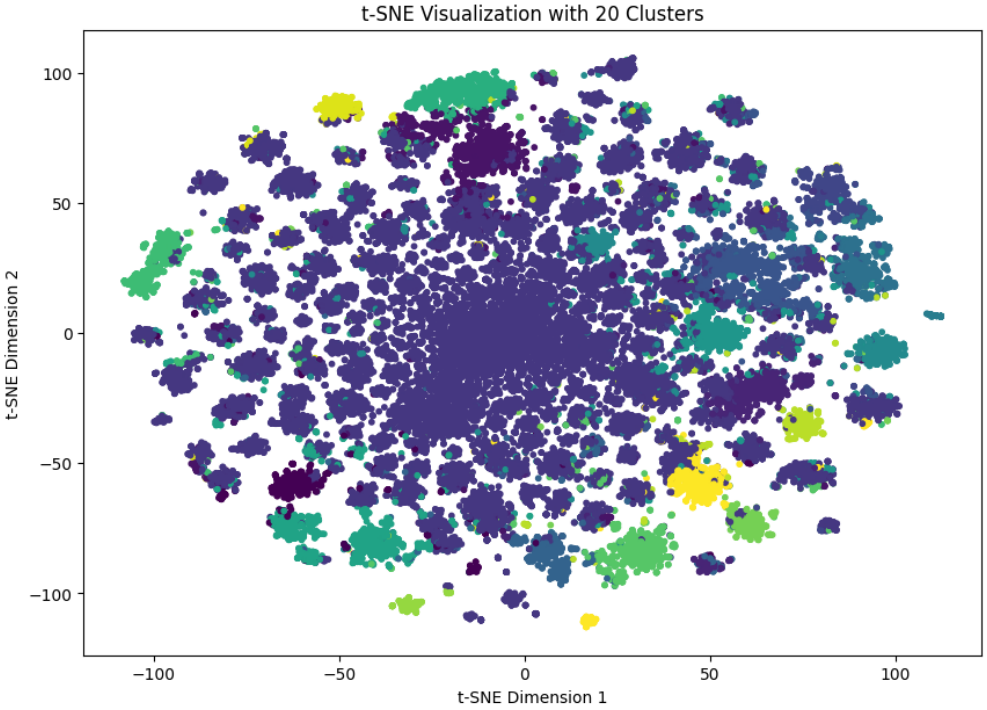

Figure 12: K-means Clustering with k = 20 and t-SNE Visualization

```
User ID:  U13740
Top Cluster:  3
What user have read so far:
['N34694', 'N42782', 'N45794', 'N55189', 'N18445', 'N63302', 'N31801', 'N19347', 'N10414']

Recommendation for user:
N42620: Heidi Klum's 2019 Halloween Costume Transformation Is Mind-Blowing   But, Like, What Is It?
N31801: Joe Biden reportedly denied Communion at a South Carolina church because of his stance on abortion
N55189: 'Wheel Of Fortune' Guest Delivers Hilarious, Off The Rails Introduction
N43142: Former NBA first-round pick Jim Farmer arrested in sex sting operation
N29177: Miguel Cervantes' Wife Reveals Daughter, 3, 'Died in My Arms' After Entering Hospice Care
N16715: Mitch McConnell snubbed by Elijah Cummings' pallbearer in handshake line at U.S. Capitol ceremony
N18870: Here Are the Biggest Deals We're Anticipating for Black Friday
N55743: 17 photos that show the ugly truth of living in a tiny house
N52551: Pamela Anderson gets backlash after wearing a Native American headdress for Halloween
N61864: The News In Cartoons
```

Figure 13: K-means Recommendation Result

```
User ID: U13740
Top 10 Recommendations:
--------------------------------------------------------------------------------
News ID: N63832
Hybrid Score: 0.5501
Content Score: 0.9042
Collaborative Score: 0.1960
Reason: Based on similar content you liked.
Title: Celtics' Gordon Hayward to have surgery on fractured hand
Abstract: Boston Celtics forward Gordon Hayward will have surgery on his left hand Monday, according to Adrian Wojnarowski.
--------------------------------------------------------------------------------
News ID: N16732
Hybrid Score: 0.5431
Content Score: 0.9110
Collaborative Score: 0.1753
Reason: Based on similar content you liked.
Title: John Brannen will 're-evaluate' status of Cincinnati basketball star Jarron Cumberland
Abstract: Cincinnati basketball star Jarron Cumberland will not play Thursday against Alabama A&M for an undisclosed reason.
--------------------------------------------------------------------------------
News ID: N13077
Hybrid Score: 0.5392
Content Score: 0.9031
Collaborative Score: 0.1753
Reason: Based on similar content you liked.
Title: High school football playoffs Round 1 scores in Cincinnati and Kentucky, Round 3 in Indiana
Abstract: Follow along live for scores and updates from Greater Cincinnati, Northern Kentucky and Southeastern Indiana.
--------------------------------------------------------------------------------
News ID: N13990
Hybrid Score: 0.5355
Content Score: 0.8958
Collaborative Score: 0.1753
Reason: Based on similar content you liked.
Title: Walmart employee accused of taking $50k from cash drawers
Abstract: nan
--------------------------------------------------------------------------------
News ID: N36569
Hybrid Score: 0.5349
Content Score: 0.8946
Collaborative Score: 0.1753
Reason: Based on similar content you liked.
Title: 31 Simple Self-Care Ideas to Get You Started
Abstract: A person shares a list of self-care activities they have used in the past and encourages readers to find what works for them!
--------------------------------------------------------------------------------
```

Figure 14: BERT hybrid Recommendation Results

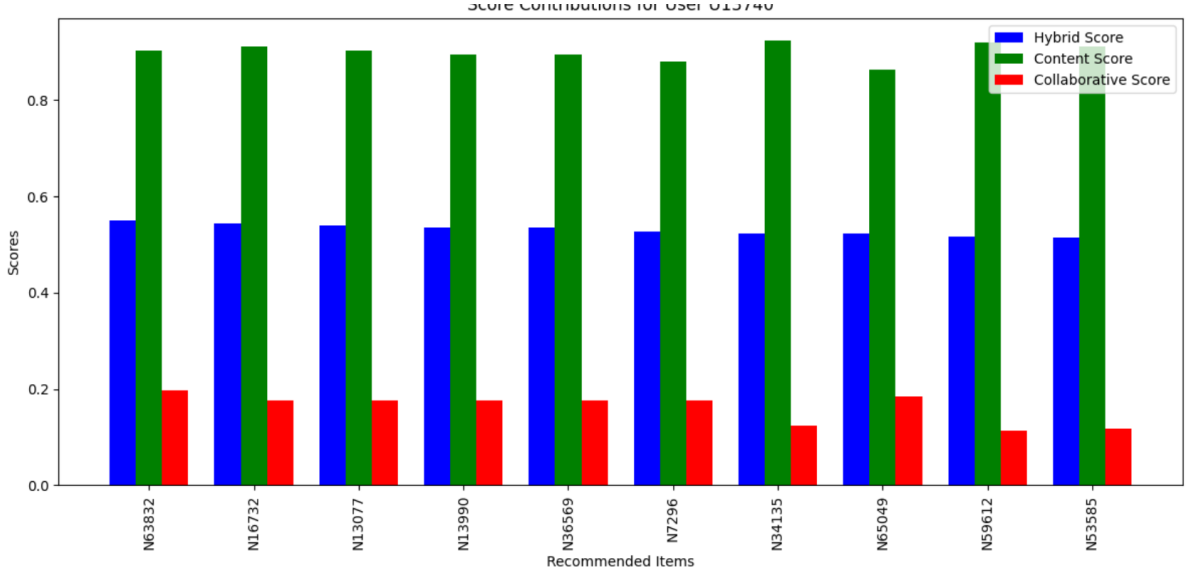

Figure 15: BERT hybrid Recommendation Scores Comparison

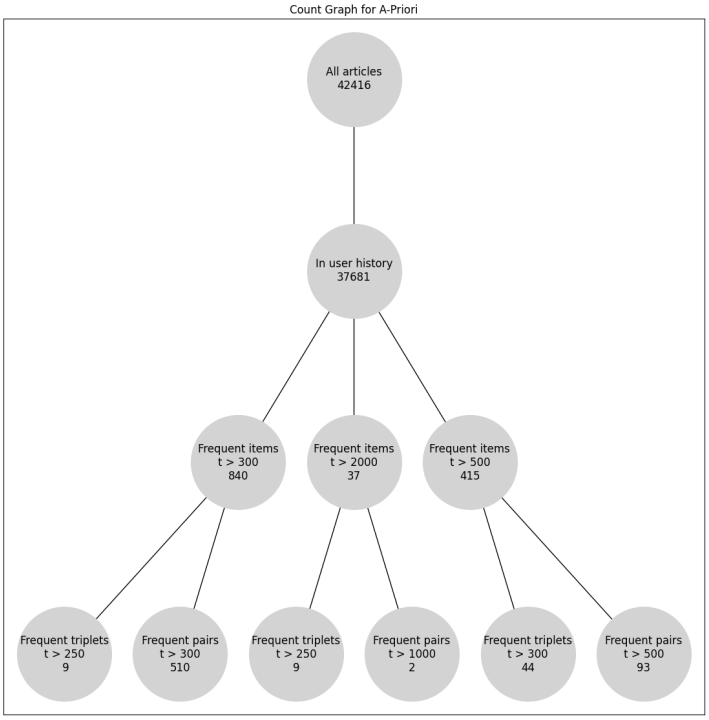

Figure 16: A-Priori Algorithm with different thresholds

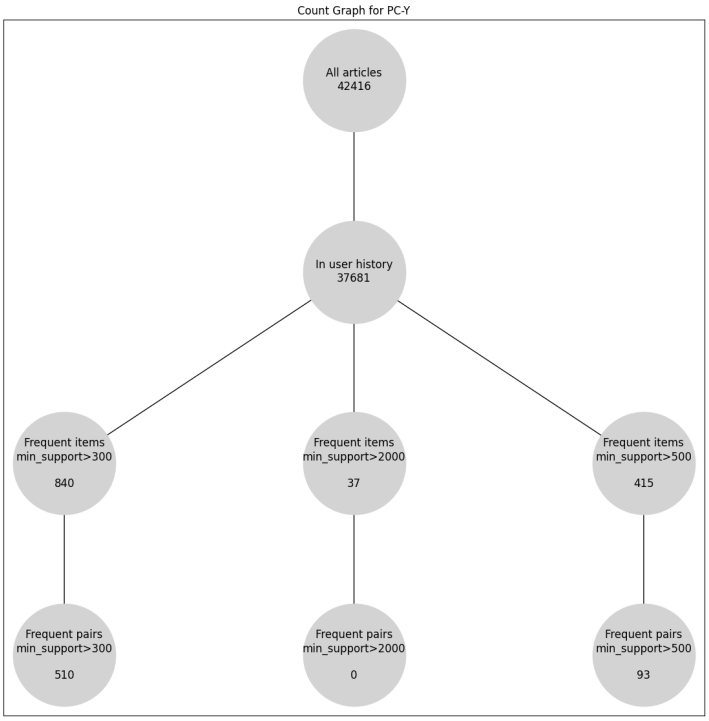

Figure 17: PCY Algorithm with different thresholds

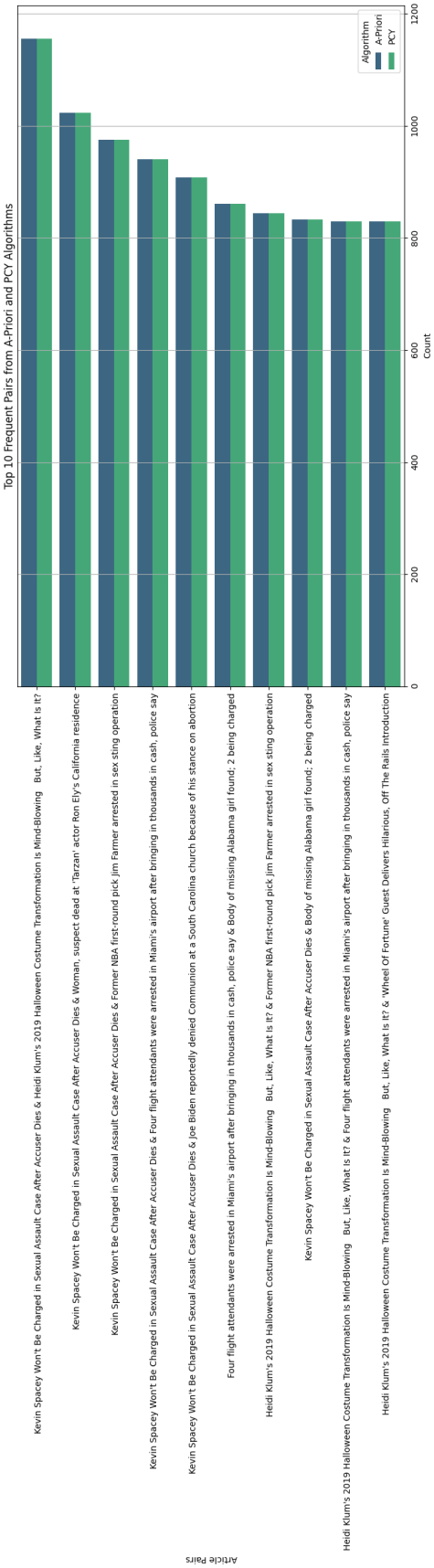

Figure 18: Top 10 Frequent Articles from A-Priori and PCY

```
Recommended Articles:
Article ID: N104737, Title: Kevin Spacey Won't Be Charged in Sexual Assault Case After Accuser Dies, Reason: Recommended
based on frequent pair with input article.
Article ID: N91597, Title: Heidi Klum's 2019 Halloween Costume Transformation Is Mind-Blowing   But, Like, What Is It?,
Reason: Recommended based on candidate pair with input article.
Article ID: N72571, Title: Former NFL lineman Justin Bannan arrested for attempted murder, Reason: Recommended based on
candidate pair with input article.
Article ID: N85484, Title: NFL world reacts to officials handing Packers win over Lions, Reason: Recommended as a frequent
article in the same category.
Article ID: N112324, Title: Boxer Patrick Day dies after suffering traumatic brain injury in super welterweight fight,
Reason: Recommended as a frequent article in the same category.
Article ID: N42718, Title: What Tom Brady, Lamar Jackson Told Each Other After Patriots-Ravens, Reason: Recommended as a
frequent article in the same category.
Article ID: N66666, Title: MLB bans women who flashed their chests behind home plate during Game 5 of World Series,
Reason: Recommended as a frequent article in the same category.
Article ID: N46994, Title: Frustrated Antonio Brown has active morning on Twitter, Reason: Recommended as a frequent
article in the same category.
Article ID: N106403, Title: NFL Week 9 Power Rankings: More ammo for Belichick as greatest coach ever, Reason: Recommended
as a frequent article in the same category.
Article ID: N44431, Title: NFL Week 8 Power Rankings: Old-school football rules the day, Reason: Recommended as a frequent
article in the same category.
```

Figure 19: Frequent Items Recommendation Results for Input Article N71977

```
Recommended articles for user 32204:
Article ID: N64332, Similarity: 0.2802, Title: Archbishop Joseph Kurtz updates his condition, says surgery was successful
Article ID: N54597, Similarity: 0.2755, Title: First look at 2020 Toyko Olympic course, which is ready for the Games
Article ID: N52355, Similarity: 0.2572, Title: Cincinnati Zoo pairs up cute rescue dog Remus and baby cheetah Kris for a 'BFF sleepover'
Article ID: N46578, Similarity: 0.2488, Title: Celtics' Tacko Fall, Tremont Waters Reenact Manute Bol-Muggsy Bogues Photos
Article ID: N44767, Similarity: 0.2445, Title: Nashville police arrest fifth suspect in Maplewood High student death
```

Figure 20: DQR Recommendations and Scores

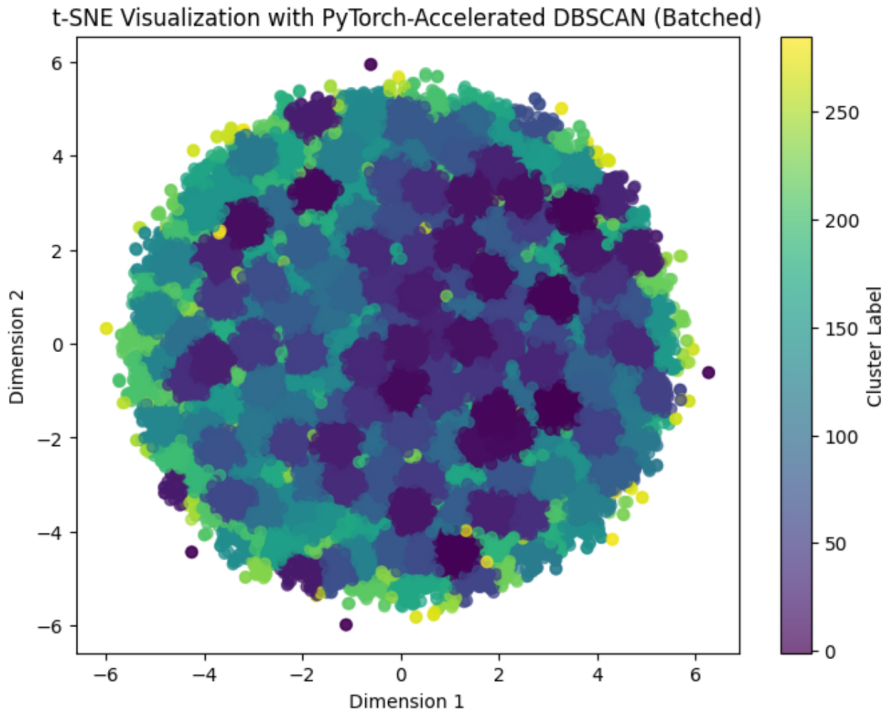

Figure 21: DQR Cluster with DBSCAN

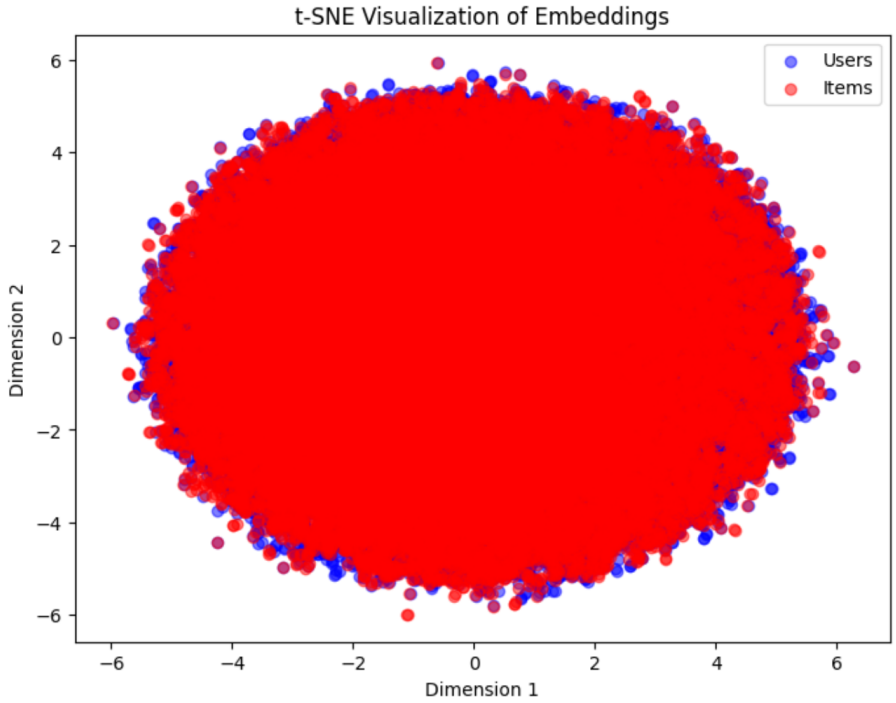

Figure 22: DQR Embeddings of User and Articles

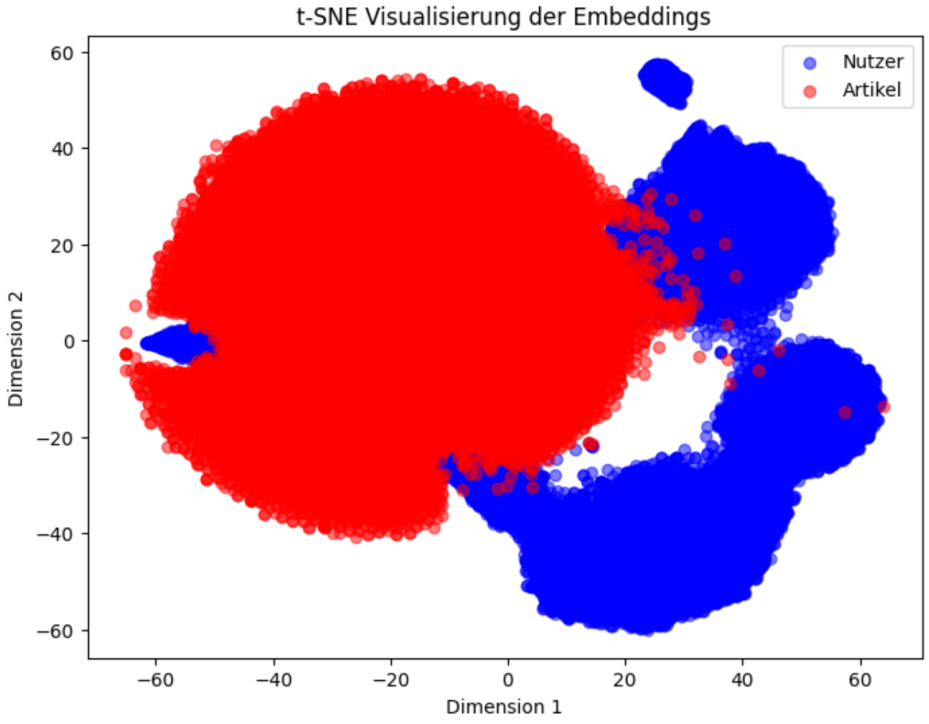

Figure 23: DQR Embeddings of User and Articles with a much more complex Agent

Group Member Contributions

As required by the formal guidelines to ensure transparency and fairness in the assessment process, we hereby confirm that all group members contributed actively and fairly to the various sections of this report. Should further clarification be needed, we are available to provide additional details.

Link to GitHub Repository

Here is the link to our notebooks for the project: https://github.com/Jus-tiine/CTDS