Fragrance Clusters: Identifying Common Perfume Groups Using Scent Profiles and Recommendation System

Christian Joshua Alberto
College of Computing and Information Technologies
National University - Manila
Manila, Philippines
albertocq@students.national-u.edu.ph.

Mark Rhey Anthony De Luna
College of Computing and Information Technologies
National University - Manila
Manila, Philippines
delunamd@students.national-u.edu.ph

Rodney Lei Estrada
College of Computing and Information Technologies
National University - Manila
Manila, Philippines
estradard@students.national-u.edu.ph

***Abstract*—Choosing a fragrance is an nuanced and personal endeavor, which can be shaped by variations in scent profiles and individual taste. The emergence of the perfume classification system The dirt-pot with a scent: A machine learning approach to cluster perfumes into groups based on scent. Due to its publicly available nature, the dataset consists of perfume names, brands, and fragrance notes. This data was preprocessed, applying Term Frequency-Inverse Document Frequency (TF-IDF) to vectorize scent descriptions and Truncated Singular Value Decomposition (SVD) to reduce dimensions. We then used the K-Means clustering algorithm to cluster together similar fragrances, with the ideal number of clusters found through the Elbow Method and the Silhouette Score. Cosine similarity was also used to improve perfume recommendations based on scent profile similarities. The findings show that this method successfully classifies perfumes to meaningfully fragrance groups and allows accurately recommendation of new perfumes. These findings provide important insights for both consumers and the fragrance industry, and allow for a data-driven and personalized approach to perfume discovery.**

I. Introduction

The art of perfumery has evolved significantly, with a vast choice of fragrances available to cater to diverse fragrance preferences. Traditional classifications, such as floral, woody, or oriental, provide a foundational understanding but often lack the granularity required for personalized experiences. The subjective nature of scent perception further complicates the development of a universally accepted classification system. In response, recent advancements have leveraged machine learning and data-driven methodologies to enhance fragrance classification and recommendation systems.

Fragrance selection is inherently subjective and influenced by factors such as personal preference. Conventional recommendation approaches, such as relying on expert opinions or customer reviews, often fail to capture the nuanced relationships between different scent compositions. Machine learning (ML) offers a robust solution by systematically analyzing fragrance compositions and consumer preferences to provide data-driven recommendations.

Several recent studies have demonstrated the effectiveness of Machine Learning in scent classification and recommendation. For example, research on machine learning for odor perception modeling has shown that algorithms can predict odor intensity and similarity with high accuracy by analyzing chemical compositions [9]. Furthermore, Natural Language Processing (NLP) techniques, such as Sentence-BERT, have been employed to convert perfume descriptions into semantically meaningful embeddings, enabling more precise fragrance recommendations [10].

This study aims to build upon these advancements by identifying common perfume groups through the clustering of fragrances based on their scent profiles. Employing machine learning techniques, specifically K-Means clustering, the research seeks to group perfumes with similar olfactory characteristics. Subsequently, a recommendation system utilizing cosine similarity will be developed to assist users in discovering fragrances that align with their personal preferences. By providing a structured clustering model and an intuitive recommendation system, this research has the potential to significantly improve the consumer experience within the fragrance industry by enabling more personalized and accurate fragrance recommendations.

II. Review of Related Literature

1) Overview of Key Concepts and Background Information

The study of fragrance classification and recommendation has evolved with advancements in machine learning and data-driven techniques. Traditional fragrance classification relies on broad categories such as floral, woody, or oriental, but these groupings often lack the specificity required for personalized recommendations.

Machine learning provides a structured approach to fragrance analysis, allowing for precise clustering and recommendation systems. One widely used method is Term Frequency-Inverse Document Frequency (TF-IDF), which converts text-based fragrance descriptions into numerical representations, making it easier to compare scent compositions [10]. Truncated Singular Value Decomposition (SVD) is then used for dimensionality reduction, improving the efficiency of clustering algorithms [8].

The algorithm used in this study is K-Means clustering, which groups perfumes with similar scent compositions. The Elbow Method and Silhouette Score are commonly used to determine the optimal number of clusters. Additionally, cosine similarity is applied to enhance recommendations by measuring the closeness of perfumes based on their base notes. These techniques collectively create a data-driven perfume recommendation system that aligns with user preferences.

Historical Development:
Historically, perfume classification was largely subjective, relying on expert evaluations and consumer preferences. However, the introduction of computational models has significantly refined fragrance grouping, leading to improved accuracy in scent prediction and recommendation [9].

2) **Review of Relevant Research Papers**
Relevant Studies:

- *Odor Perception Modeling Using Machine Learning*: This study demonstrated that algorithms can predict odor intensity and similarly with high accuracy by analyzing chemical compositions. These models have improved the scientific understanding of scent relationships and have been applied in the fragrance industry[9].
- *Natural Language Processing (NLP) for Perfume Recommendation*: Studies using **Sentence-BERT** and other NLP models have successfully converted textual descriptions of perfumes into semantically meaningful embeddings, enabling more refined fragrance recommendations [10].
- *Sentiment Driven Community Detection in a Network of Perfume Preferences*: This study utilized community detection techniques on a perfume co-preference network derived from user reviews to cluster similar perfumes [12]. The authors enhanced sentiment analysis and improved the modularity of detected communities. but focuses on user preferences expressed through reviews rather than inherent scent profiles. The use of sentiment analysis and network analysis offers valuable insights into user behavior, which could complement the current study's approach.

Relation to Current Study:

These studies both inform and contrast with the current research. The odor perception and NLP studies provide valuable context for prior work in fragrance analysis. The community detection study offers a complementary perspective, focusing on user preferences and sentiment rather than scent profiles. While the HF prediction study explores different ML models, its findings on model tuning and data preprocessing are directly applicable to the current research's methodology. Unlike prior studies that focus on ensemble models or user reviews, this research investigates the potential of K-Means clustering and cosine similarity applied to scent profiles to achieve accurate and personalized fragrance recommendations.

3) **Prior Attempts to Solve the Same Problem**
Notable contributions to fragrance classification and recommendation include research on odor perception modeling using chemical compositions [9], the application of NLP techniques like Sentence-BERT for fragrance description analysis [10], and the use of TF-IDF and SVD for feature extraction in fragrance recommendation systems [8, 10]. Many fragrance companies also utilize proprietary recommendation systems, though details are often undisclosed.

Successes and Shortcomings:
Previous work has successfully leveraged data-driven approaches to understand and categorize fragrances. Odor perception models have shown promise in predicting objective scent properties [9], while NLP techniques have enabled the capture of semantic relationships from fragrance descriptions [10]. However, these methods often struggle with the subjective nature of fragrance preference and the complexity of scent perception. Existing industry solutions may lack transparency and personalization, often relying on broad categories or purchase history. Furthermore, methods using TF-IDF and SVD, while effective for feature extraction, may benefit from more sophisticated clustering algorithms and similarity metrics. These limitations motivate the need for more nuanced and personalized fragrance recommendation systems, like the K-Means clustering and cosine similarity-based approach explored in this study.

Summary of Key Questions

1.) **What research has already been done on this topic?** Odor perception modeling [9], NLP for fragrance recommendation [10], TF-IDF/SVD for feature extraction [8, 10], community detection in perfume networks [12].

2.) **How do existing studies relate to or differ from your research?** Relates by using similar techniques (TF-IDF, SVD). Differs by focusing on scent profiles, using K-Means and cosine similarity.

3.) **What are the key methods, theories, and models that inform your work?** TF-IDF, SVD, K-Means clustering, Elbow Method, Silhouette Score, cosine similarity.

4.) **What gaps exist in the current literature that your research fills?** Focus on scent profiles, combined approach (TF-IDF, SVD, K-Means, cosine similarity), personalized recommendations.

5.) **How does your proposed approach contribute to advancing the field?** More nuanced recommendations, data-driven approach, improved consumer experience.

6.) **What challenges or limitations do previous approaches have, and how do you address them?** Subjectivity of scent (using scent profiles), complexity of scent perception (structured scent profiles), lack of transparency/personalization (clear methodology, personalized recommendations), limitations of TF-IDF/SVD (potential for future exploration of other techniques).

7.) **How does your research align with or challenge current trends?** Aligns with the trend of using ML in the fragrance industry. Challenges by offering an alternative to user-review based recommendations, focusing on inherent fragrance properties

III. Methodology

A. Data Collection

The dataset used in this study was sourced from a publicly available repository containing detailed information about perfumes. The dataset comprises attributes such as perfume name, brand, fragrance notes, and user ratings. The data was originally collected by fragrance researchers and made available through online repositories. The dataset was gathered through standardized assessments, ensuring consistency in feature measurements [1].

B. Data Pre-Processing

Prior to clustering, the dataset underwent several preprocessing steps:

- **Handling Missing Data**: Missing values were identified and addressed based on feature relevance.
- **Feature Selection**: Only relevant columns such as name, brand, and notes were retained.
- **Text Normalization**: The name and brand columns were converted to lowercase to maintain uniformity.
- **Column Note Preprocessing**: Transforming and layouting the contents of Notes Column, removing unnecessary notes (values that are not perfume notes).
- **TF-IDF Vectorization**: Fragrance notes were vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) to transform textual descriptions into numerical representations suitable for clustering.

C. Experimental Setup

To ensure efficient dependency management and reproducibility, the researchers used Python Poetry to manage and maintain the project environment. Poetry streamlines the installation of libraries and handles project dependencies, version control, and virtual environments effectively

The following tools and libraries were utilized to implement, train, and evaluate the K-Means model for Fragrance Clusters Project:

- **Python**: used for the entire analysis and model development due to its versatility.
- **NumPy**: essential for numerical computations, such as matrix operations, required in model training and data transformation.
- **Pandas**: employed for data manipulation and analysis, including
- **Scikit-learn**: implemented the K-Means clustering model and provided tools for TF-IDF vectorization, dimensionality reduction, and model evaluation.
- **Matplotlib** and **Seaborn**: used for visualizing data distributions, cluster groupings, and PCA projections.

Hyperparameter tuning was conducted to determine the optimal number of clusters using the Elbow Method and Silhouette Score

D. Algorithm Selection

K-Means. clustering approach while it also identified natural structures in fragrance information. Numerous clustering systems depend on the K-Means algorithm because it combines practicality with successful results for matching similar components through numerical representations [3]. With K-Means clustering the research team managed to classify perfumes across different groups while identifying fragrance patterns.

K-Means functions through first running k centroids followed by assigning data points to their nearest centroid and continuously recalculating centroids until the system converges.

The objective function is:

$$J = \sum_{j=1}^{K} \sum_{n \in \mathcal{S}_j} \left| x_n - \mu_j \right|^2,$$

figure 1. *K-Means Algorithm*

where *xn* represents a data point, μ*j* represents the centroid of cluster Ci, and the sum of squared distances ensures that points are as close to their respective centroids as possible

E. ***Cosine Similarity.*** analysis measured both the similarity levels between different fragrances using the TF-IDF vectorized fragrance notes. The equivalence measurement of textual data through two vector dimensions occurs via the cosine function that determines the angle's measure. Vectors retain greater importance than their length in text-based analysis since their orientations affect similarity measurements according to [4].

The formula for cosine similarity measurement appears as follows:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

figure 2. *Cosine Similarity Algorithm*

The calculation uses two vectors representing dot product while and stand for their individual magnitudes. Application of cosine similarity in the model allowed the grouping of perfumes with shared fragrance notes regardless of different textual descriptions which led to better cluster coherence.

Truncated Singular Value Decomposition (SVD). uses essential dataset features through a method which decreases computational requirements. The decomposing mechanism of SVD generates three output matrices from an initial matrix.

$$A = U\Sigma V^T$$

figure 3. *Truncated SVD Algorithm*

The SVD transformation consists of orthogonal matrices *U* and *V* the diagonal matrix . This matrix contains the singular values. The preservation (sigma sign) of significant data structure occurs after truncating because the technique maintains critical components even while reducing dimensionality.

It shows SVD succeeds as a pre-processing technique for high-dimensional data because it decreases noise elements while sustaining key features for better clustering results [8]. This research successfully applied SVD to display clustering results which generated improved understanding about how fragrances relate to their themes. The reduction of dimensions achieved through SVD made the clustering algorithm function efficiently without compromising vital information needed for meaningful fragrance categorization

F. *Training Procedure*

1. Load and preprocess dataset
2. Apply TF-IDF vectorization on fragrance notes
3. Determine optimal cluster count using silhouette analysis
4. Train K-Means model and assign perfumes to clusters
5. Perform SVD for dimensionality reduction and visualization
6. Analyze Clusters to identify common top notes

G. *Evaluation Metrics*

The assessment of clustering methods included these evaluation metrics for evaluation purposes:

- **Silhouette score** evaluates the cluster separation and compactness while a higher score indicates optimal cluster definition.
- **Top Notes Consistency** evaluates the clustering methods' ability to group frequently occurring scent notes correctly.
- **Value-based assessment** method checks whether expensive perfumes correctly match clusters for premium

The performance evaluation of the recommendation system uses clustering results to measure their effectiveness in generating custom recommendations between users and their preferred fragrance notes.

IV. Results and Discussions

4.1. *Silhouette Score Elbow Method for Optimal Clustering.* The Silhouette Score operated as the main measure to establish the most fitting cluster number. The Silhouette Score determines cluster quality through a dual analysis of how tightly clusters fit together with each other and how distinct their separation should be.

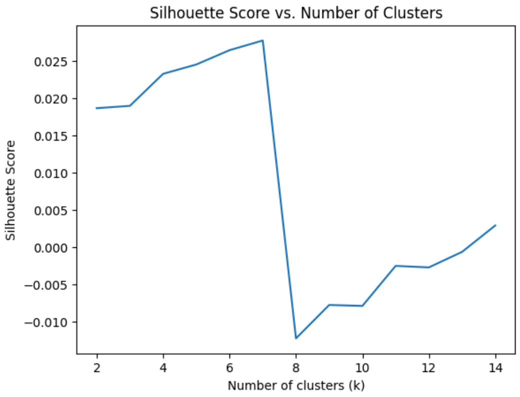

figure 4. *Silhouette Score vs. Number of Cluster*

When k=7 reached its highest value which demonstrated the groups created at this configuration were optimally separated. Both Silhouette Score and Elbow Method selected k=7 as the ideal number of clusters since higher values produced no additional performance enhancement. The application of Silhouette Score validated the meaningful nature of clustering because the method produced well-separated and compact clusters.

4.2. *Truncated Singular Value Decomposition (SVD).* Truncated Singular Value Decomposition (SVD) was applied to reduce dimensionality and improve the interpretability of the clusters. The first two principal components were visualized to observe how well the perfumes were separated.

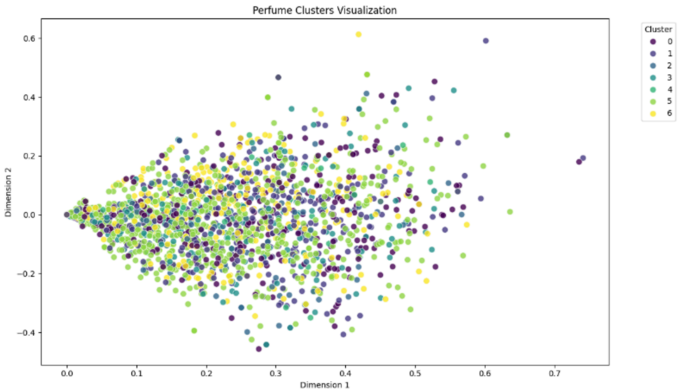

figure 5. *Truncated SVD Chart*

SVD successfully reduced data complexity while retaining essential fragrance features. The first two principal components captured a significant percentage of variance, reinforcing the effectiveness of clustering. However, the visualization of clusters revealed substantial overlapping, making it difficult to see distinct groups clearly. One primary reason for this overlap is the high-dimensional nature of fragrance data. When reducing the dataset to only two dimensions, some cluster-specific information is inevitably lost, causing perfumes from different clusters to appear closer than they actually are in the full feature space. Additionally, some perfumes share similar fragrance profiles across multiple clusters, contributing to soft boundaries rather than distinct separations.

4.3. *Top Notes Analysis.* To further validate clustering quality, the dominant top notes in each cluster::

- **Cluster 0 (Fresh & Floral Woody Musk)** was characterized by citrus-based fragrances, including Musk, Amber, Bergamot, Rose, Jasmine, Sandalwood which are commonly associated with fresh & Floral Woody Balance

  Cluster 0 Top Terms: musk, amber, bergamot, rose, jasmine, sandalwood, lemon, cedar, vanilla, violet

- **Cluster 1 (Oriental & Spicy Woody))**was dominated by floral scents such as Patchouli, Saffron, Oud, Amber, Leather, Cardamon, typically linked to Deep, rich, and exotic fragrances, often found in Middle Eastern perfumery.

  Cluster 1 Top Terms: patchouli, saffron, rose, sandalwood, oud, amber, cedar, cinnamon, leather, cardamom

- **Cluster 2 (Earthy & Citrus Woody)** comprises woody and musky fragrances like Cedarwood, Vetiver, Bergamot, Nutmeg, Grapefruit , Fresh, woody, and slightly spicy colognes, often masculine and grounding often perceived as sophisticated and long-lasting.

  Cluster 2 Top Terms: cedarwood, vetiver, bergamot, musk, patchouli, sandalwood, nutmeg, cardamom, lemon, grapefruit

- **Cluster 3 (Smoky & Resinous Oriental)** was defined by spicy and oriental notes such as Incense, Patchouli, Amber, Leather, Vetiver, Vanilla perceives dark, resinous, smoky fragrances, ideal for colder seasons which evoke warmth and exotic appeal.

  Cluster 3 Top Terms: incense, patchouli, amber, bergamot, musk, leather, cedar, sandalwood, vetiver, vanilla

- **Cluster 4 (Fruity Floral Musk)** predominantly contains fruity and sweet notes, including Peony, Lychee, Musk, Vanilla, Rose, Freesia which are playful, sweet, and airy fragrances, often in feminine scents.

  Cluster 4 Top Terms: peony, musk, vanilla, lychee, bergamot, sandalwood, rose, violet, peach, freesia

- **Cluster 5 (Classic Green & Woody Floral)** featured aquatic and fresh notes, such as Vetiver, Jasmine, Cedar, Oakmoss, Lemon, Iris, often associated with Earthy, natural, and vintage-style perfumes with chypre elements.

  Cluster 5 Top Terms: vetiver, bergamot, jasmine, sandalwood, cedar, musk, lemon, iris, rose, oakmoss

- **Cluster 6** encompasses gourmand and creamy scents like Vanilla, Sandalwood, Jasmine, Patchouli, Tuberose which add smooth, sweet, and seductive scents with floral-woody warmth.

  Cluster 6 Top Terms: vanilla, sandalwood, jasmine, bergamot, patchouli, cedar, musk, violet, rose, tuberose

4.4 *Insights.* The clustering method functioned to organize perfumes through their fragrance profiles which produced an organized system for fragrance classification. The cluster segregation in perfume data remained obscure because of its high dimensions coupled with its sparse characteristics. Research demonstrates the need to select suitable methods of dimensionality reduction. Due to its complex nature, fragrance data needs non-linear reduction techniques like t-SNE and UMAP to achieve better visualization results.

Visually the lack of clean cluster boundaries indicates some fragrant characteristics fail to produce marked differences between groups. The performance of fragrance clusters requires additional contextual details including perfume intensity measurements as well as longevity estimates and seasonal factors to achieve better results. The inability to visualize clusters reveals the limitations of working with limited data since several perfumes have insufficient recorded characteristics to achieve distinct cluster groupings.

V. Conclusions

In this research, we used machine learning techniques to overcome the problem of fragrance classification and recommendation. Conventional perfume houses classify scents only by broadly labeled categories, like floral or woody, making personalized suggestions difficult. The purpose of this research was to classify perfumes according to their scent profile using K-Means clustering and enhance recommendations using cosine similarity.

The study was able to successfully labelled perfumes into seven groups based on its fragrance using TF-IDF vectorization, dimension reduction via Truncated SVD, and K-Means clustering. The ideal number of clusters was found by the Elbow Method and the Silhouette Score. Results showed that perfumes containing similar fragrance compositions were grouped together in an accurate manner, thereby improving the efficiency of scent-based recommendations.

The work presented is a structured, data-based perfume classification and recommend method. This research extends categorical classification of perfumes by joining text-based fragrance definitions and clustering algorithms. Similar scent profiles leads to personalized recommendations using cosine similarity.

These shifts carry new implications for consumers and the fragrance industry alike. This model makes it easier for consumers to discover which perfumes they would prefer to buy, and fragrance brands and retailers can use this model to optimize marketing channels and further engage consumers. This work also showcases the applicability of machine learning methods in subjective fields, exemplified here by perfumery.

Although the clustering method adequately delineated fragrance clusters, the results were influenced by some limitations. Due to the high-dimensional nature of the data for fragrance, there was some overlap between the clusters, leading to less clear classification. Compounds like perfume longevity, seasonal factors, and intensity weren't included either, which could lead to tighter models in the future.

Other futures works should consider more complex machine learning techniques such as deep learning models or hybrid ones, where clustering techniques can be used in tandem with neural networks, to improve the classification of fragrances. The recommendation system could become more personalized and drive up engagements by incorporating user preference data including purchase history and real-world scent preferences. Also, if we include external factors such as the weather, occasion or cultural preferences, one would create a complete fragrance recommendation system. Techniques like UMAP or t-SNE for dimensionality reduction could be used to improve visualization and separate clusters better. Finally, the work could be advanced by increasing the data and covering a larger set of perfume compositions and user reviews, which will help reach a more accurate and applicable model for real-world cases.

This proves the efficiency of machine learning in putting together and recommending fragrances based on the scent profiles. It lays the foundation for innovative and more data-driven perfume suggestions, by mechanizing fragrance classification techniques, which adds value to both buyers and experts in the field. Note that this letter and design have to be read and viewed in the context of very latest advances in this area at the intersection of AI and the olfactory world; future developments might offer even greater impact in this space.

References

[1] Perfume Recommendation Dataset. (2021, February 10). Kaggle. https://www.kaggle.com/datasets/nandini1999/perfume-recommendation-dataset

[2] Heng, Y. P., Lee, H. Y., Chong, J. W., Tan, R. R., Aviso, K. B., & Chemmangattuvalappil, N. G. (2022). Incorporating machine learning in Computer-Aided molecular design for fragrance molecules. *Processes, 10(9),* 1767. https://doi.org/10.3390/pr10091767

[3] IBM (2024, December 19) K-Means Clustering. *What is k-means clustering?* https://www.ibm.com/think/topics/k-means-clustering

[4] Jain, A (2024, August 5). TF-IDF Vectorization with Cosine Similarity - Anurag Jain - Medium. *Medium*. https://medium.com/%40anurag-jain/tf-idf-vectorization-with-cosine-similarity-eca3386d4423

[5] Hanafizadeh, P., Zareravasan, A., & Khaki, H. R. (2010). An expert system for perfume section using artificial neural network. *Expert Systems with Applications, 37(12),* 8879-8887. https://doi.org/10.1016/j.eswa.2010.06.008

[6] Sushma, S., Sundaram, N., & Jayapandian, N. (2021). Machine learning - based unique perfume flavour creation using quantitative structure-activity relationship (QSAR). *Proceedings of the International Conference on Computer, Communication, and Material Science (ICCMC).* https://doi.org/10.1109/ICCMC51019.2021.9418246

[7] Rodrigues, B. C. L., Santana, V., V., Murins S., & Nogueira, I. B. R. (2024, February 19) *Molecule generation and optimization for efficient fragrance creation.* arXiv.org. https://arxiv.org/abs/2402.12134

[8] Pramoditha, R. (2023, June 27). Truncated SVD for dimensionality reduction in sparse feature matrices. *Medium*. https://rukshanpramoditha.medium.com/truncated-svd-for-dimensionality-reduction-in-sparse-feature-matrices-c083b4af7ddc

[9] Rugard, M., Jaylet, T., Taboureau, O., Tromelin, A., & Audouze, K. (2021). Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages. *PLoS ONE*, *16*(5), e0252486. https://doi.org/10.1371/journal.pone.0252486

[10] Kim, J., Oh, K., & Oh, B. (2024). An NLP-Based perfume note estimation based on descriptive sentences. *Applied Sciences*, *14*(20), 9293. https://doi.org/10.3390/app14209293

[11] Rodrigues, B. C. L., Santana, V. V., Queiroz, L. P., Rebello, C. M., & Nogueira, I. B. R. (2024). Scents of AI: Harnessing graph neural networks to craft fragrances based on consumer feedback. *LSRE-LCM, Faculty of Engineering, University of Porto & Chemical Engineering Department, NTNU.*

*[12]*Kalashi, K., Saed, S., & Teimourpour, B. *(2024, October 24). Sentiment-Driven community detection in a network of perfume preferences. arXiv.org. https://arxiv.org/abs/2410.19177*