1 Method's Current Usage

1.1 Combine Clustering and Frequent Itemsets Mining to Enhance Biomedical Text Summarization

Authors: Oussama Rouane, Hacene Belhadef, Mustapha Bouakkaz
Published: Expert Systems With Applications, 2019
Summary: This study combines K-means clustering and Apriori-based frequent itemset mining for biomedical text summarization. Sentences are clustered based on semantic similarity, and frequent itemsets are mined within each cluster to identify important patterns. Sentences covering the most significant itemsets are selected for summarization, ensuring global and subtopic-specific information is retained.
Takeaway: This project demonstrates how to apply frequent item analysis to filter important sets for simplification, inspiring my simplification of genre.

1.2 Contextual Market Basket Analysis during Covid-19

Authors: Gisela Christy Mooy, Sani M Isa
Published: Journal of Social Science, 2023
Summary: This study uses Apriori-based Market Basket Analysis (MBA) to uncover shifts in purchasing patterns before and during Covid-19. Transactions were categorized by seasons (e.g., holidays, ordinary days), and frequent itemsets were mined to analyze changes in consumer behavior. For example, the focus shifted from beauty products to essential items like germ cleaners during the pandemic.
Takeaway: This paper demonstrates how frequent item analysis can adapt to contextual changes, inspiring analysis of genre associations over time.

1.3 Discovering HIV-Related Information by Means of Association Rules and Machine Learning

Authors: Lourdes Araujo, Juan Martinez-Romo, Otilia Bisbal, Ricardo Sanchez-de-Madariaga
Published: Scientific Reports, 2022
Summary: This study introduces a semi-supervised method, EXTRAE, combining frequent pattern growth (FP-Growth) and machine learning to extract relevant association rules (ARs) for identifying relationships between HIV-related diseases. By starting with a small labeled seed dataset and iteratively refining with unsupervised learning, the method effectively filters meaningful ARs while minimizing irrelevant ones. Key findings include relationships confirmed by medical experts and plausible new links for further investigation.
Takeaway: This study showcases how frequent item analysis can be paired with semi-

supervised learning, to identify significant patterns in complex datasets. This advanced idea might be useful for future explorations.

2 My Experimentation

2.1 Goal

I used *FPGrowth* in PySpark to perform **frequent item analysis** to explore patterns among music genres, for example, which genre has the highest number of tracks? Which genres are associated?
As later I plan to analyze the evolution of music by genre, but there are just too many genres. This analysis here can help me understand the connections between genres and simplify the classification of genres.

2.2 Challenges

1. **Parameterization**: FPGrowth has several parameters, like `minSupport` and `minConfidence`. Those parameters are crucial for the quality of the results, and they are sensitive to dataset. Thus customized tuning is needed.
2. **Visualization**: The origin results from FPGrowth are not intuitive, making it hard to understand the connections between genres. Therefore, I need to find a good visualization plan to make the results more understandable.

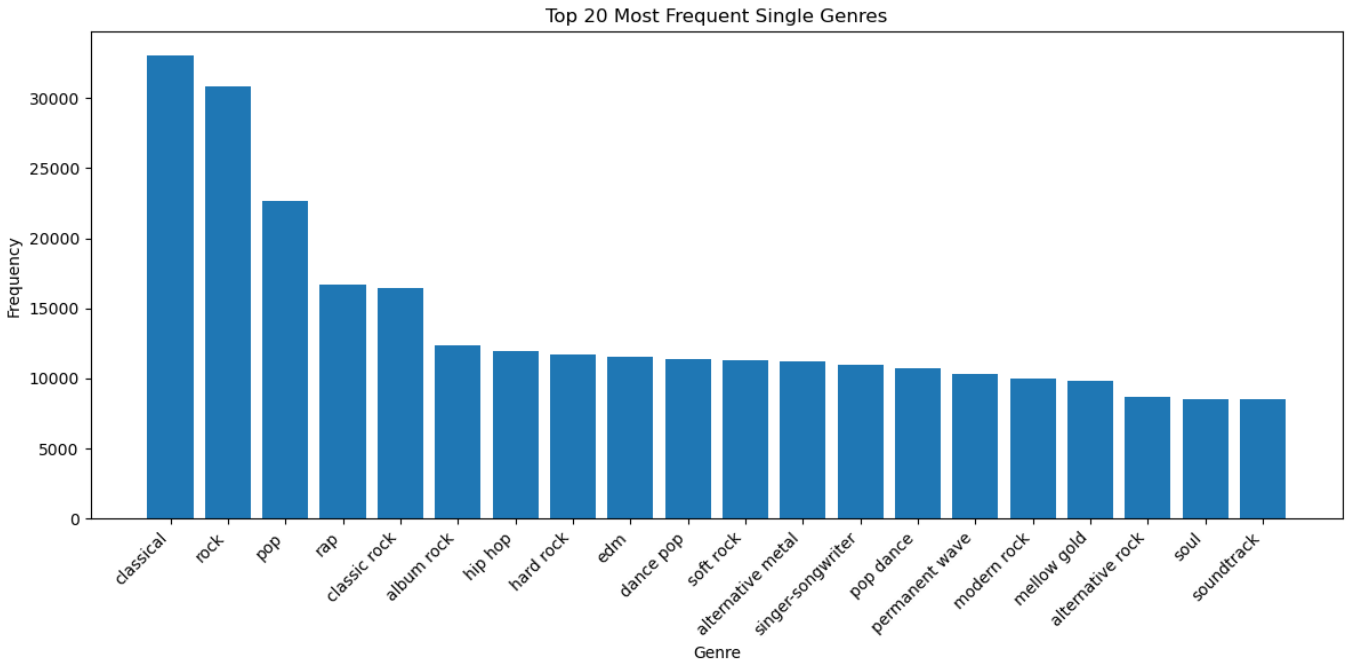
2.3 Process

1. Initially, I used default parameters without specifying `minSupport` or `minConfidence`. No item sets or association rules were generated due to overly stringent criteria. （According to the documentation FPGrowth on PySpark,by default, `minSupport` = 0.3, `minConfidence` = 0.8）
2. Then I adjusted the parameters to lower thresholds, specifically `minConfidence` = 0.40, `minSupport` = 0.005. As the criteria was relaxed, the range of item sets and association rules were expanded, revealing results that align with intuition and domain knowledge.
3. Visualization: I further visualized the results by:
   • Bar chart: the Frequency of different item sets. I have three charts, for top 20 single, pair, or triple item sets, respectively.
   • Network graph: nodes represent genres and directed edges represent discovered rules. Larger nodes represent more frequent genres. Blue edges represent strong associations.

3 Findings and Implications

3.1 Most Frequent Genres

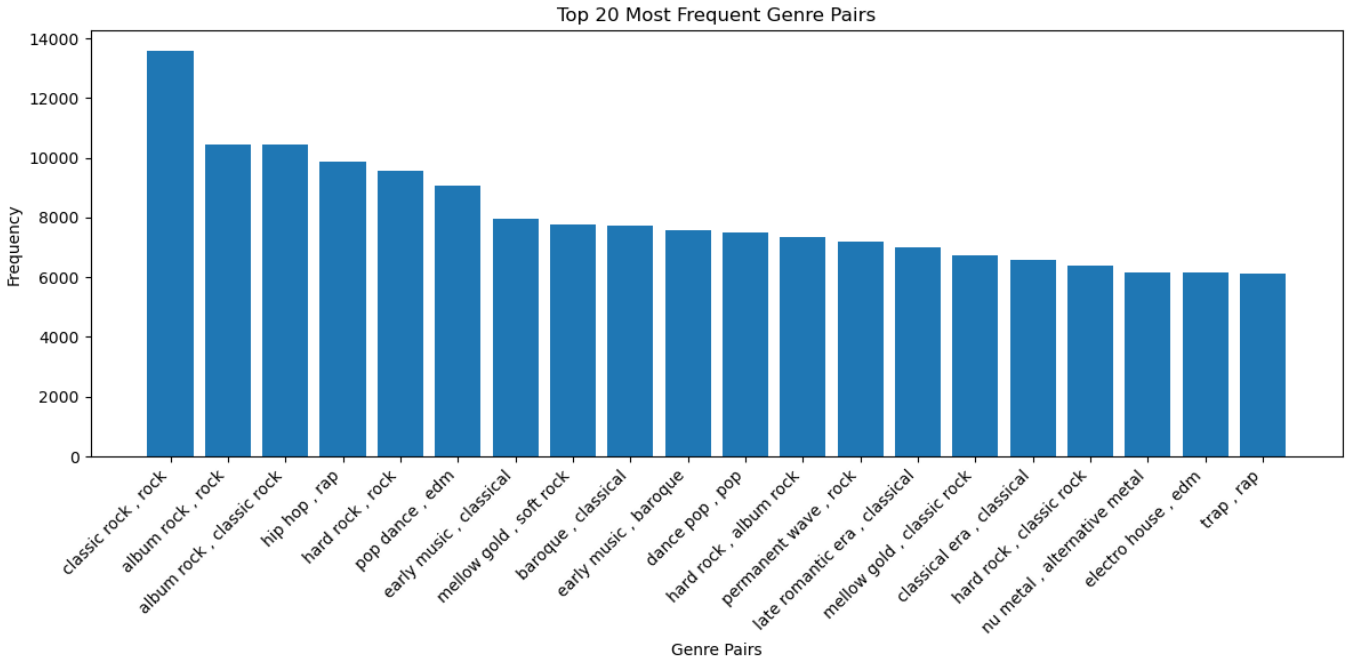
• Classical, rock, and pop are the top three most frequent genres, dominating the single-item analysis. This indicates their prominence in the global music market.



3.2 Genre Pairs

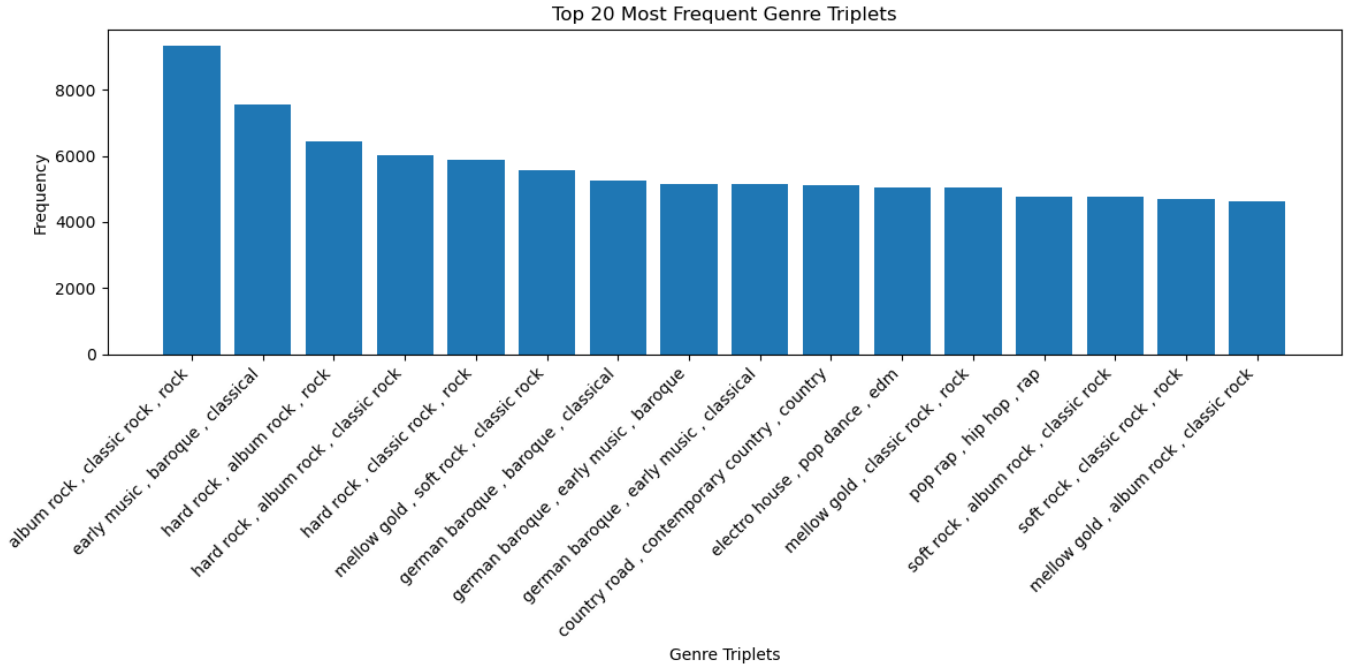
• High-frequency genre pairs, such as “classic rock - rock” and “album rock - classic rock,” demonstrate strong associations within similar categories. This implies hierarchical

relationships or sub-genre connections.



3.3 Genre Triplets

• Triplets like “album rock - classic rock - rock” and “early music - baroque - classical” imply thematic clusters or historical continuity within musical styles.

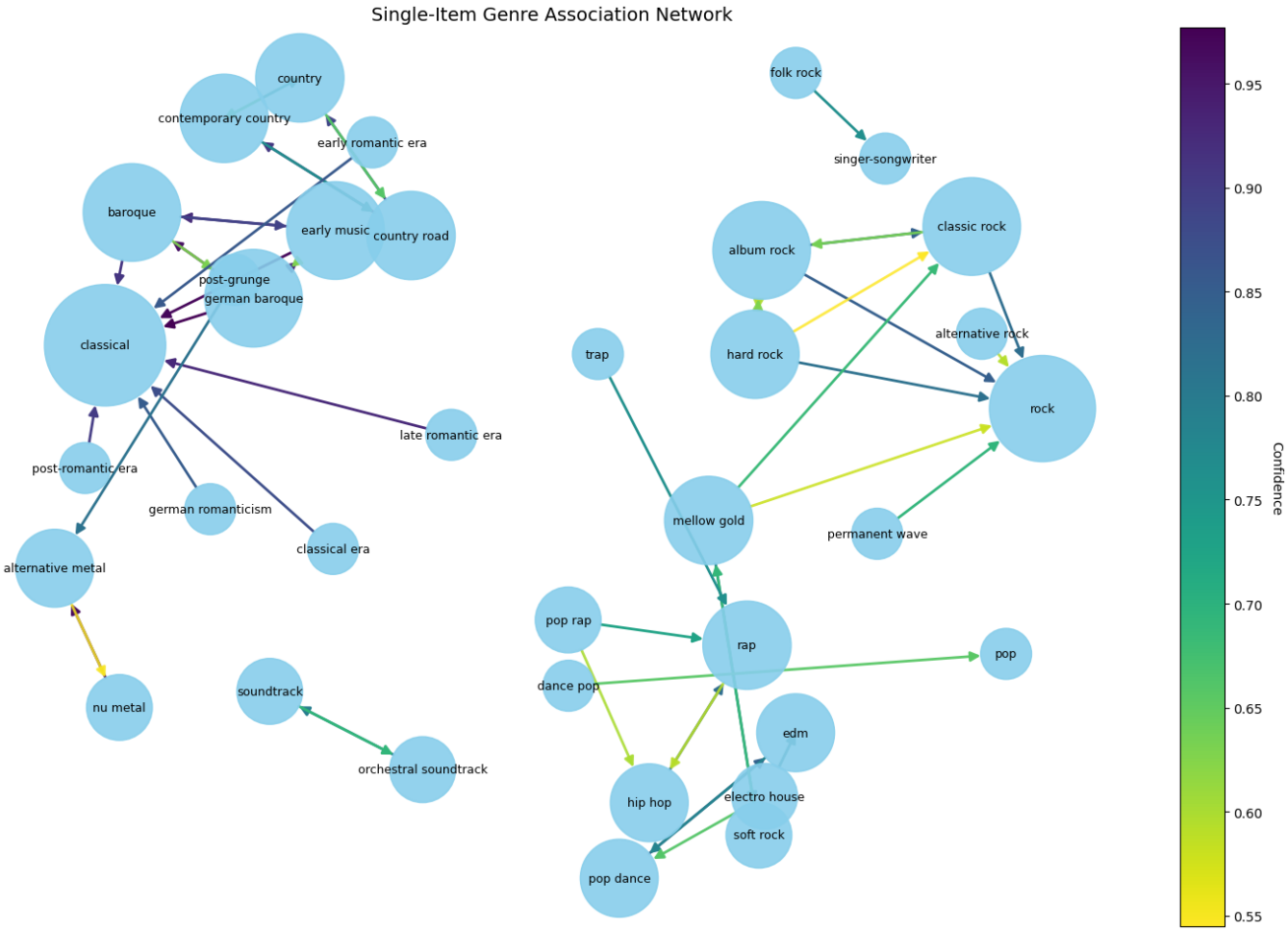


3.4 Network Graph

• The network diagram presents the connection between genres more intuitively. It reveals meaningful genre groups, such as "rock" connecting various rock sub-genres.

• We can also observe that the level of subdivision of different music genres varies. For

example, classical has more subdivision than rap, which may be explained by the longer history of classical music.



3.5 Summary

Thanks to frequent item analysis based on FPGrowth, we can efficiently mine potential genre item sets in music data, laying the foundation for subsequent genre-based evolution analysis. For example, treating commonly co-occurring genres as the same category can simplify genre classification.

On the other hand, mining association rules can help us understand the directed relationships between genres. For instance, with "album rock" as the antecedent and "rock" as the consequent, the confidence is high, but not vice versa, which may imply that "classic rock" might be a subclass of "rock". This helps us understand the subdivision of different music genres in a broader context.