PROJECT 2

UNSUPERVISED METHODS

Dolganiuc Ana

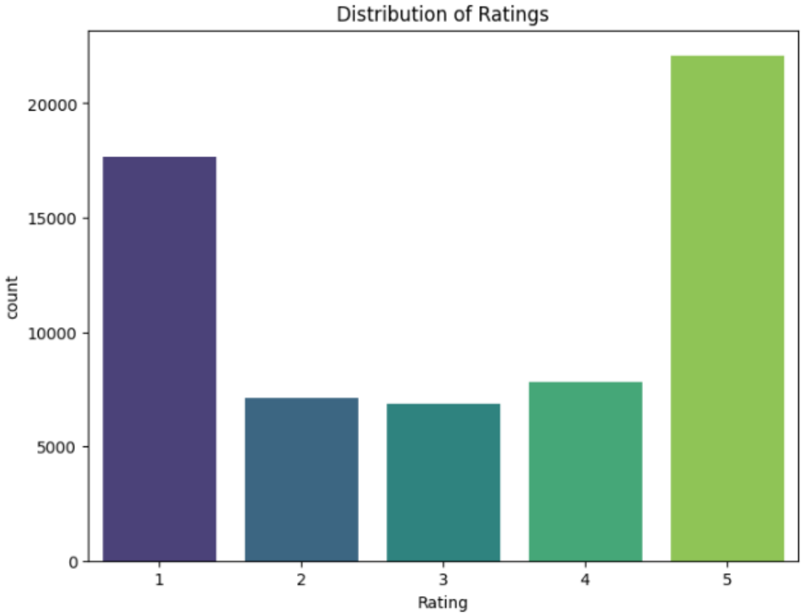
1. Purpose

In this project we will predict the clusters for reviews for the Spotify App. In order to do that, we will analyze and clean the data, do feature engineering, try to get the best accuracy using 2 unsupervised methods: GMM and LDA, and we'll compare the results with random chance and the supervised baseline predicted using the "Rating" column.

2. Exploratory Data Analysis

It has 61594 entries and no missing values.
The dataset contains 2 columns: Review and Rating

- Distribution of Ratings analysis:



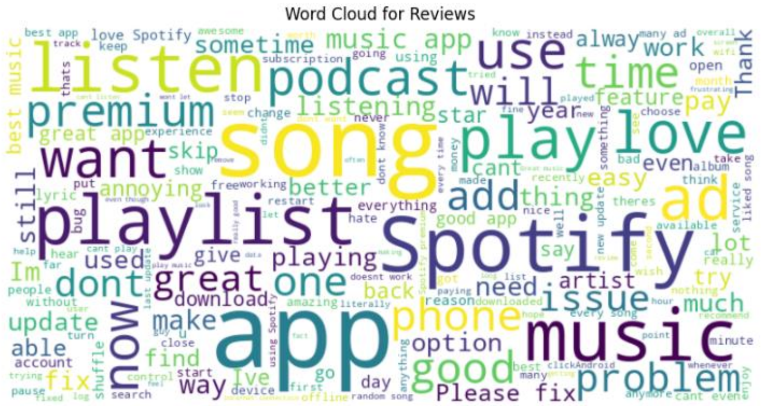
As we can see, the ratings are imbalanced, there are more reviews for the ratings 1 and 5.

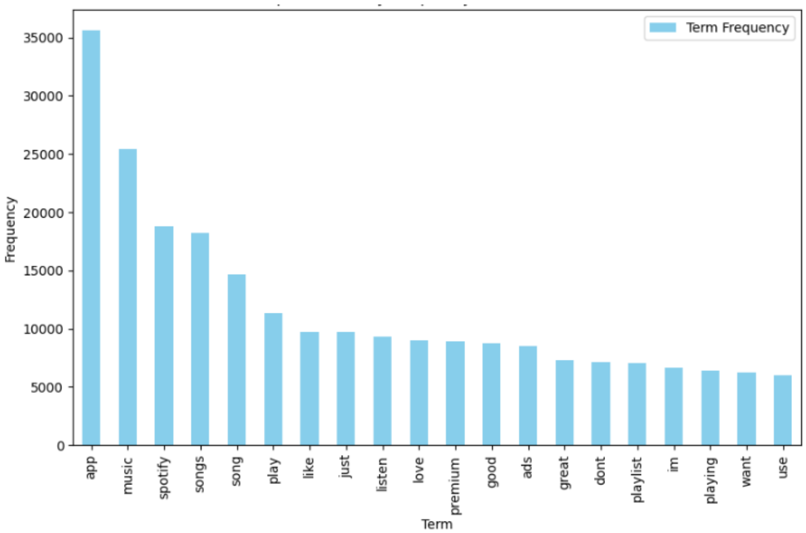
3. Data preprocessing

In order to make the reviews cleaner we'll remove all the :URLs, punctuation signs, unknown characters and numbers that don't give us too much insight.

For feature engineering we'll use:

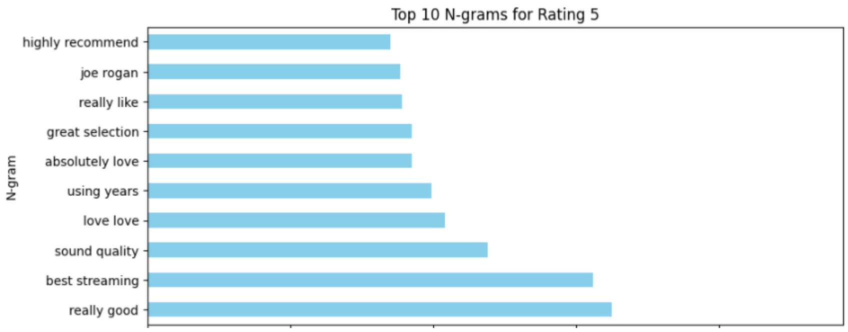
- TFIDF
- Lemmatization
- BOW with N grams
- Most common words in the reviews

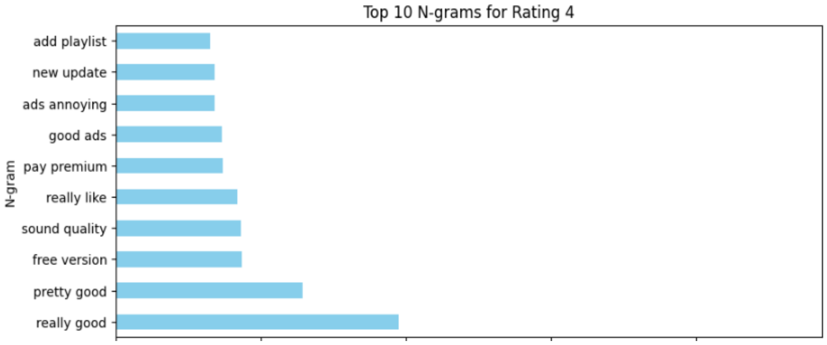


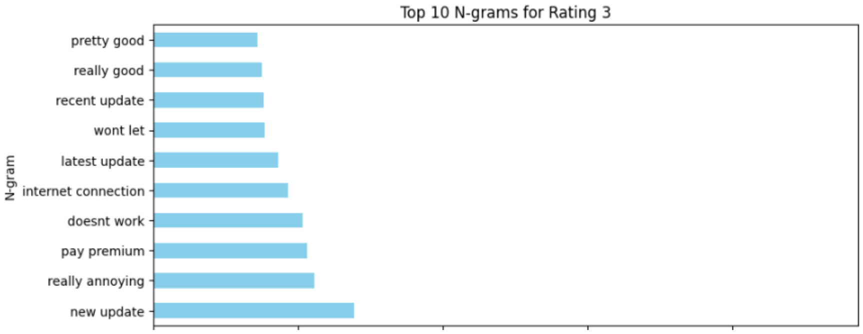


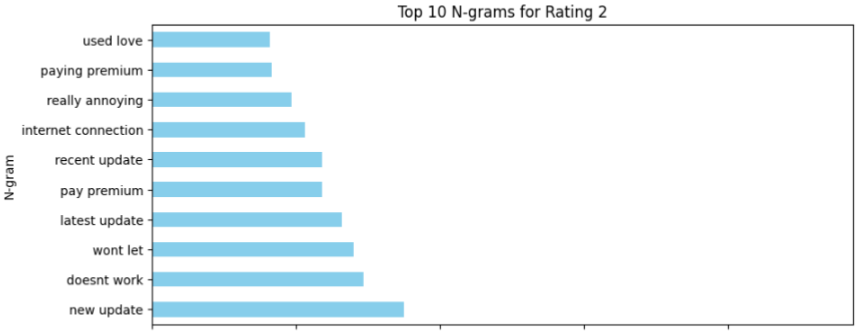
Here we can observe the most common words, because of that we will create a custom stop words list, in order to get a more accurate analysis.

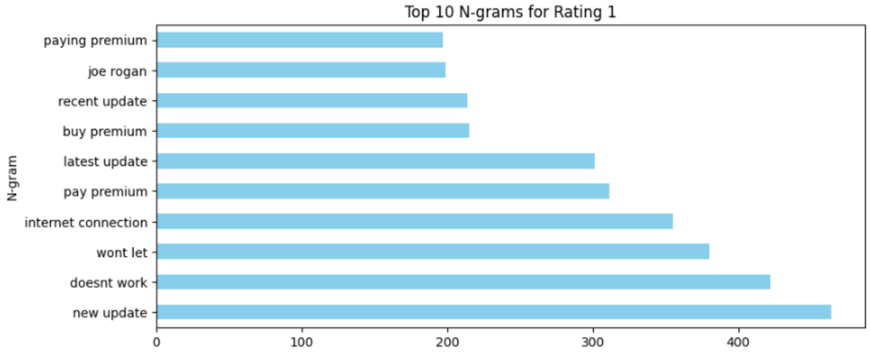
- Analyzing N-grams for every rating


Top 10 N-grams for Rating 5

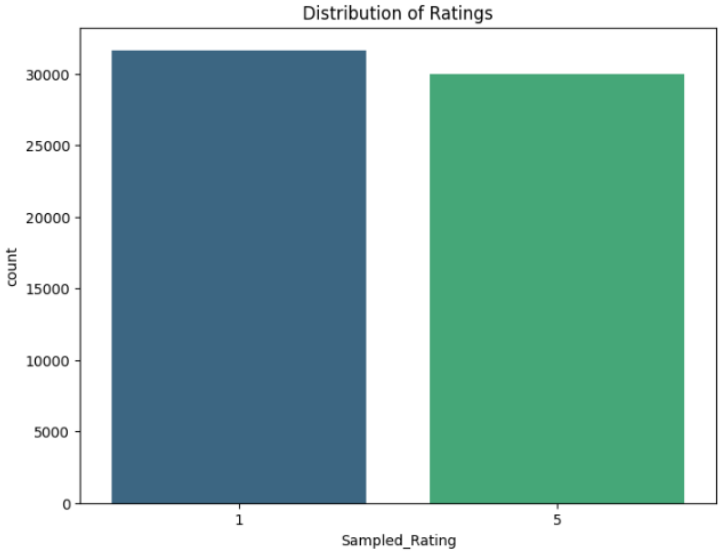

Top 10 N-grams for Rating 4


Top 10 N-grams for Rating 3


Top 10 N-grams for Rating 2

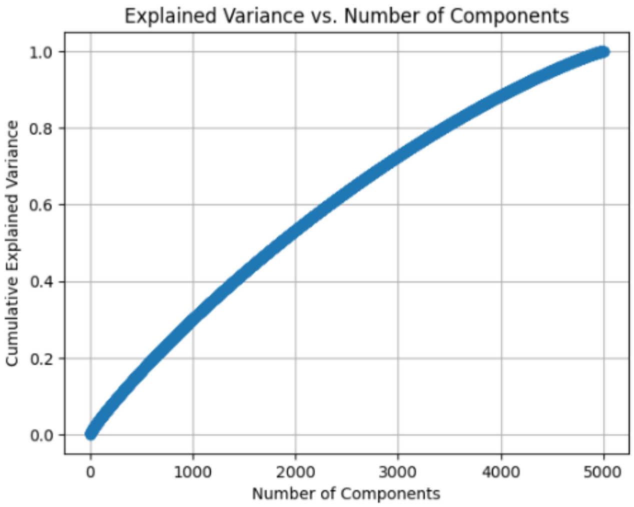


Because the one-grams didn't give us a lot of insight, we used the bi-grams to uncover patterns in the reviews. As we can see, the ratings 1,2,3 come from the new updates, problems with playing the songs and the bi-grams are similar, so we should group the ratings: 1,2,3 in a cluster and the 4,5 in a different cluster.



4. Dimensionality reduction:

We will use PCA for the GMM model, but in order to do that we need to see what number of components do we need to explain the optimal cumulative variance.

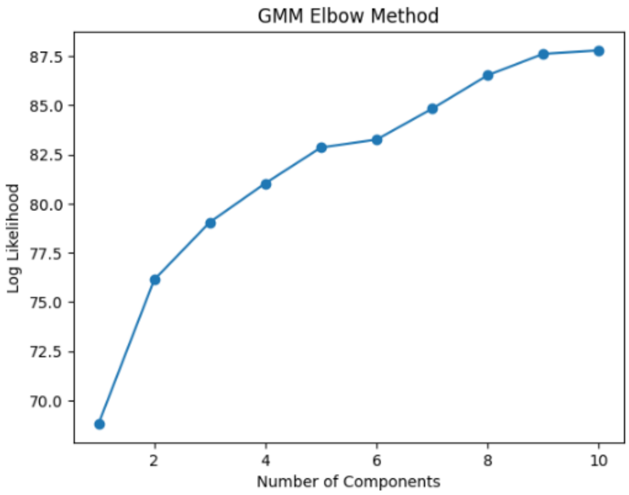


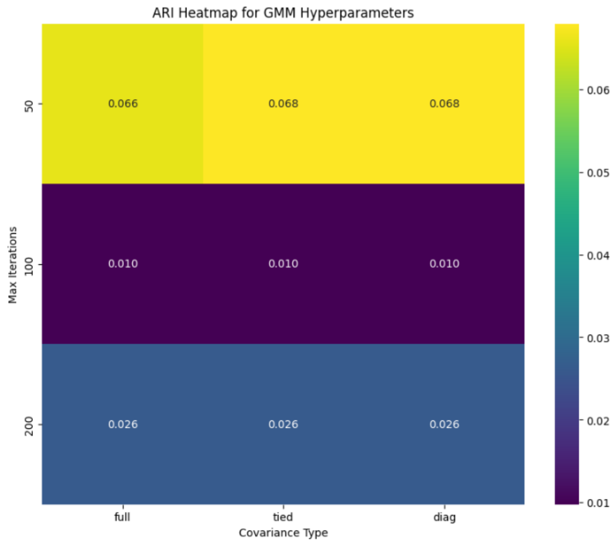
As we can see, the number of 3000 components will explain around 80% of the variance, this means that the dimensionality of the data will be reduced, while the significant patterns will be kept.

5. Unsupervised Models
    a. Gaussian Mixture Models (GMM)

GMM is a probabilistic model that models the data into clusters and each cluster is modeled as a Gaussian Distribution. Each component is characterized by its mean, covariance and weight. Unlike K-means that handle well the circular data, GMM's can handle even very oblong clusters. The convergence of the algorithm is determined by the log-likelihood that each data point belongs to each component.

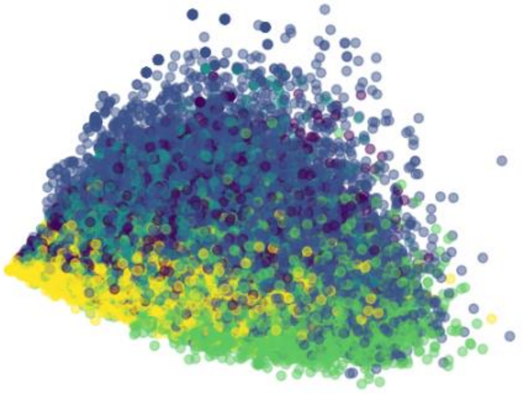
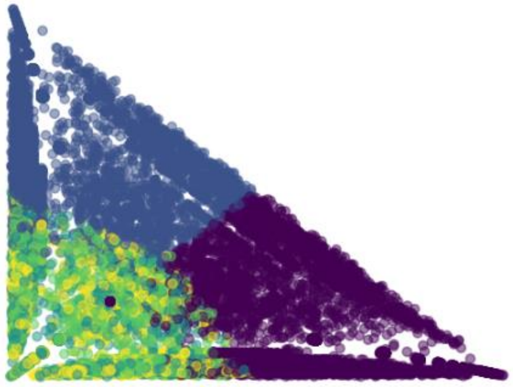
GMM hyperparameters fine tunning:





b. Latent Dirichlet Allocation (LDA)

The purpose is to assign topics to the given document and attributes each word to a certain topic. For each document it chooses the distribution of the topics and for each word in the document it chooses a topic from the distribution of topics from the document and a word from the selected topic's word distribution.

GMM clusters visualization	LDA clusters visualization
	

6. Final results

Accuracy metric: Adjusted Random Index (ARI)

Random	Logistic Regression	GMM	GMM Optimized	LDA	LDA Optimized
-8.34422	0.41120	0.08037	0.08283	0.12711	0.17927

Best Parameters LDA: {'learning_decay': 0.7, 'max_iter': 50, 'n_components': 2}

Best Parameters GMM: {'covariance_type': 'full', 'max_iter': 100, 'n_components': 5}

7. Conclusion

For this project on the unsupervised methods, I chose to cluster the reviews for the Spotify music app and compare the results with the Rating label. The process consisted of EDA where the data was cleaned and brought to a more useful form. After this I applied different preprocessing techniques like BOW, TF-IDF and Lemmatization. For the dimensionality reduction for the data I used PCA with the optimal number of components and the algorithms that I chose for training and testing the clusters where: Gaussian Mixture Models and Latent Dirichlet Allocation. Given the figures from above we can see that LDA performed way better on the set and the clusters were more defined.