Project on Big Data Visualization

Ayush Raj Aryal | Deepak Gautam

Project Goal

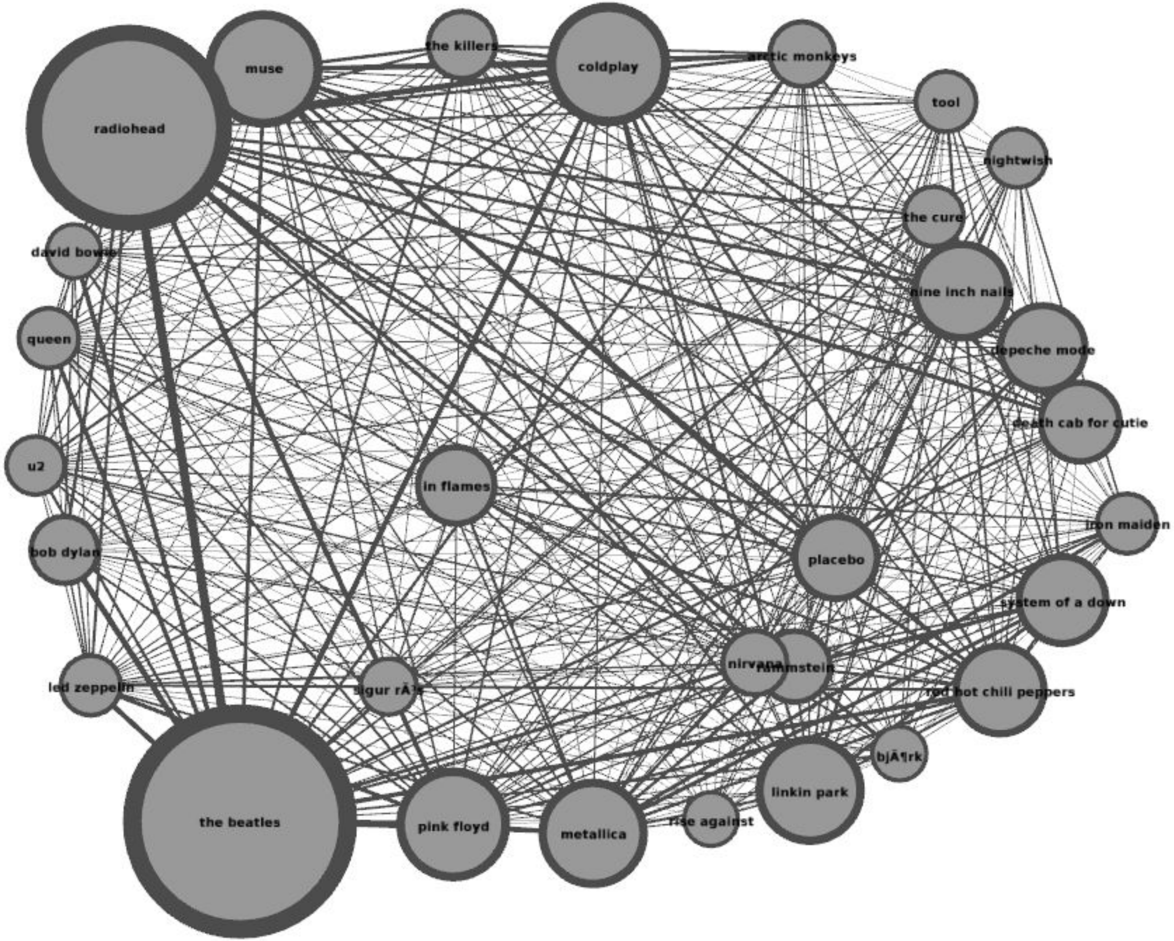
- Study on Last.fm dataset
- Community Detection and Artist Recommendation on Graphs
- Age Group Prediction based on user’s listening habits

Dataset

- Last.fm Dataset
  - Tuple of (user, artist, number of plays)
  - User Attributes: Age, Country, Gender, Date of Signup
- Total Data Size
  - 360, 000 Users (17M rows)
- Studied Sample
  - 10,000 users (~500K rows)

Graph Analysis

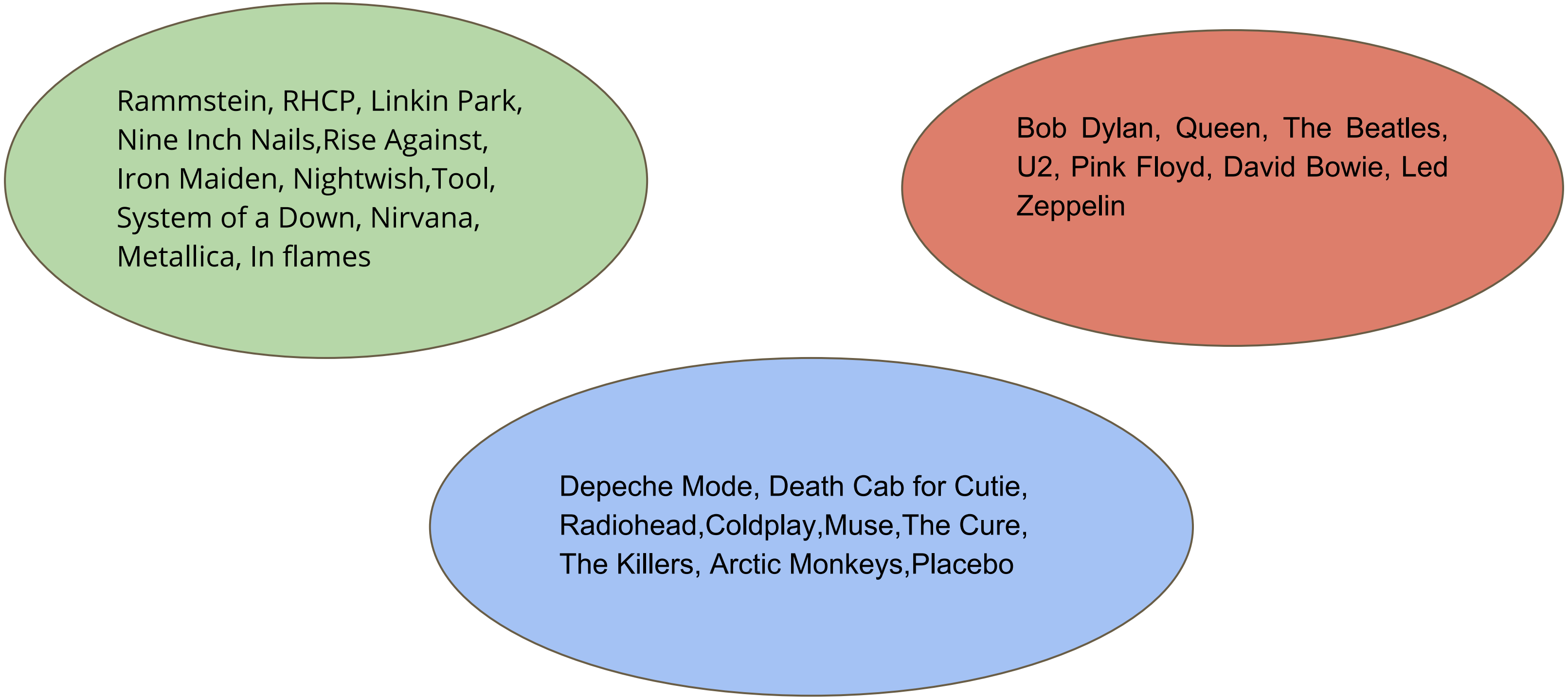
- Graph Construction
  - Python Library : Networkx
  - Artists were considered as nodes
  - Edge between artists
  - Number of plays of artists as node weights
  - Number of plays by common users between two artists as the edge weights
  - Other variants of weight calculation is also tried


radiohead
muse
the killers
coldplay
arctic monkeys
tool
nightwish
the cure
nine inch nails
depeche mode
death cab for cutie
david bowie
queen
u2
in flames
iron maiden
bob dylan
placebo
system of a down
nirvana
rammstein
led zeppelin
sigur rÃ³s
red hot chili peppers
bjÃ¶rk
linkin park
the beatles
pink floyd
metallica
rise against

Graph Analysis

- Community Detection
  - Community - A Python package that works with Networkx
    - Louvain Modularity Algorithm
  - Community Threshold - At least 3 artists needed to form a community
  - Artists seemed to be grouped by different aspects

Graph Analysis - Results

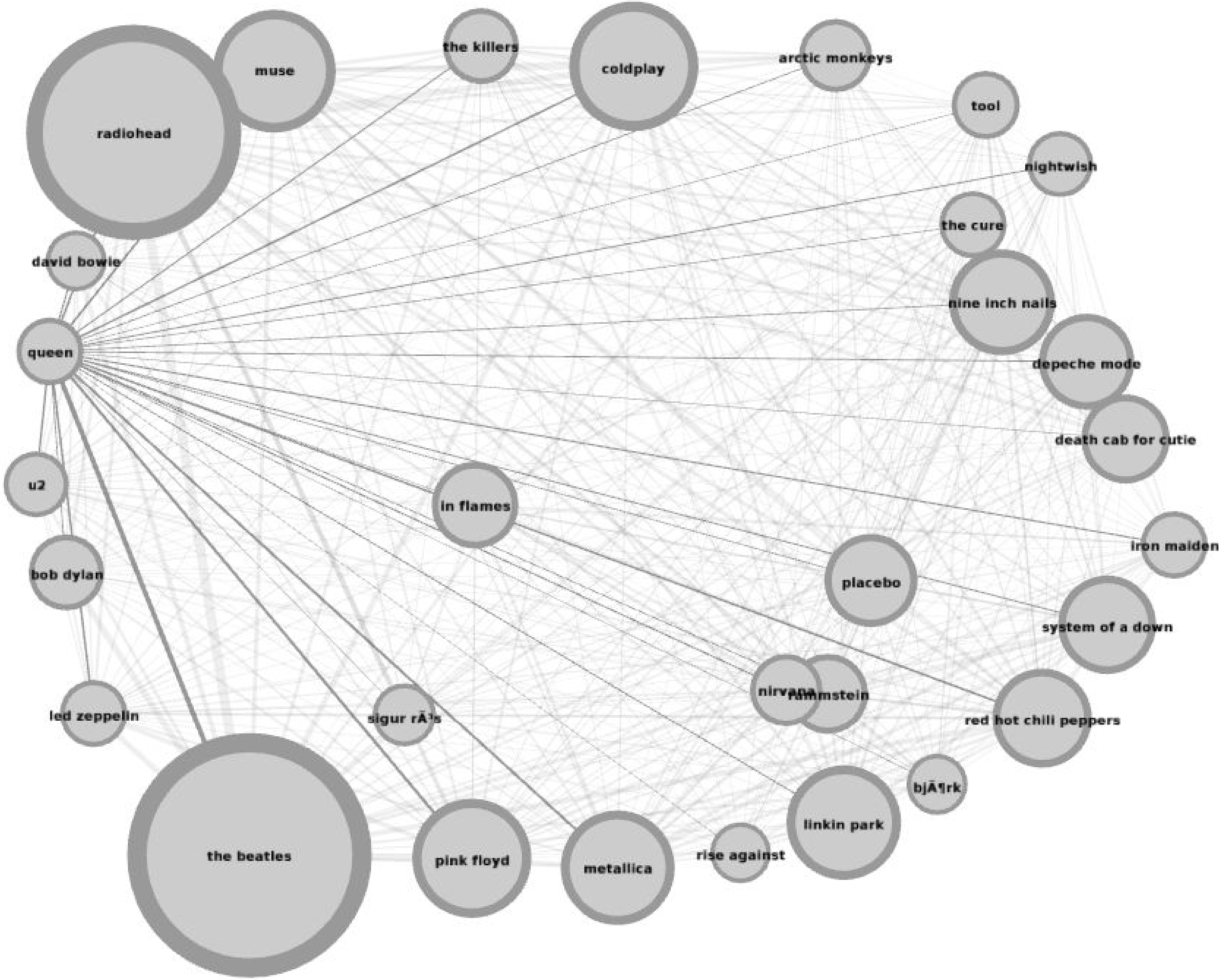


Rammstein, RHCP, Linkin Park, Nine Inch Nails,Rise Against, Iron Maiden, Nightwish,Tool, System of a Down, Nirvana, Metallica, In flames

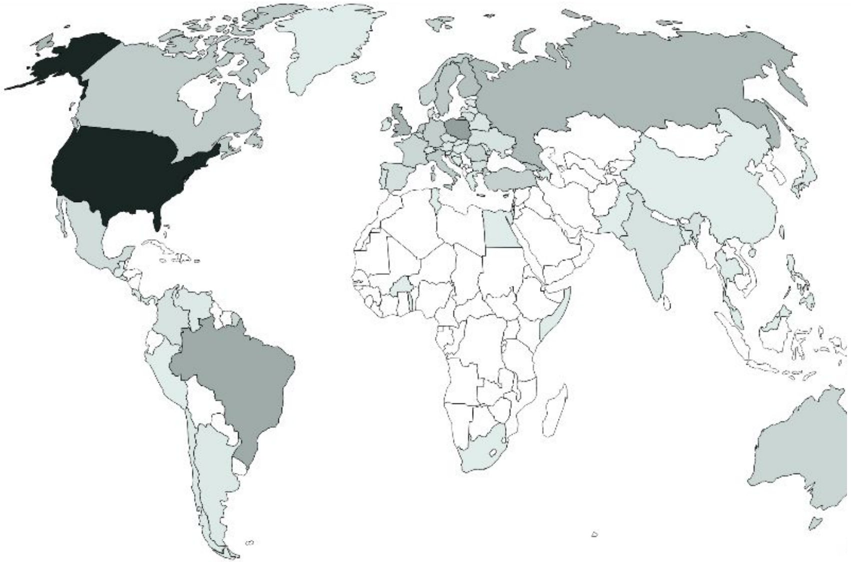
Bob Dylan, Queen, The Beatles, U2, Pink Floyd, David Bowie, Led Zeppelin

Depeche Mode, Death Cab for Cutie, Radiohead,Coldplay,Muse,The Cure, The Killers, Arctic Monkeys,Placebo

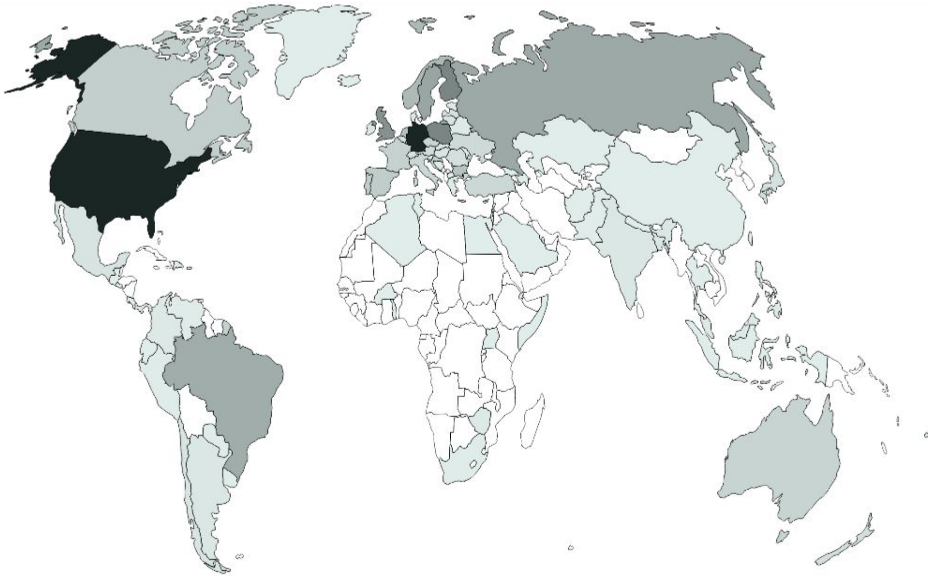
Bob Dylan, Queen, The Beatles, U2, Pink Floyd, David Bowie, Led Zeppelin



Graph Analysis- Artist vs Group



Plays of ‘Pink Floyd’



Plays of Group ‘Pink Floyd’ belongs

Graph Analysis - Talking Points

- Variation on edge weights
- Normalized weights
- Relation between grouped artists
  - Some artists are related by their genre
  - Some are related by their popularity time
- Recommendation to Users
  - Artists belonging to same group can be recommended to users

Age Prediction

- Age group based on listening habits
- Classification
- Age group as labels
- User details and their artists as features
- Scikit-Learn - Python Machine Learning Library

Data Preparation

- Noise removal
    - NaN Values
- Feature vector creation
    - Each Row represents a user
        - List of artists (String)
        - Gender (-1 or 1)
        - Country (String)
        - Date of Signup (In Months)
- Label creation
    - 0-10, 10-20, 20-40, 40+

Data Preparation - Text Vectorization

- Field to be vectorized
  - String of Artists
  - Country
- Tokenization - English Stop Words
- TF-IDF Vectorization
  - Document - String of Artists by a single user, Country
  - Corpus - String of all the artists, List of all countries

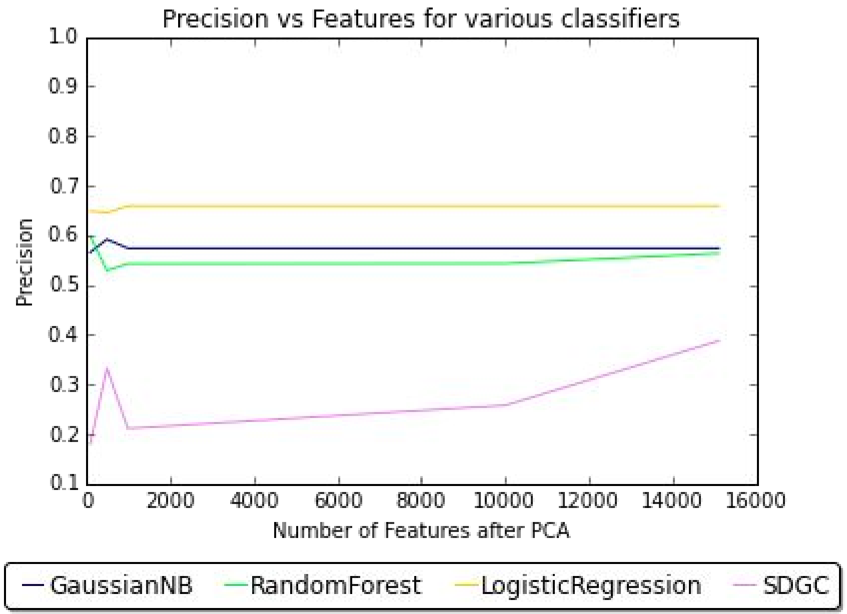
Classification

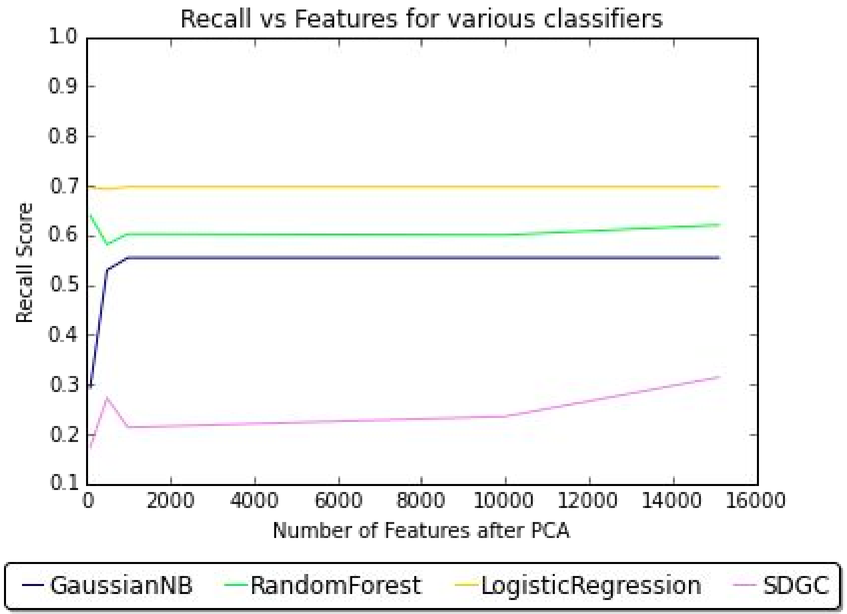
- Feature Reduction
  - PCA - changed number of components from (2-16000)
- Prediction via Classification
  - Algorithms Considered
    - Naive Bayes
    - Random Forest
    - Logistic Regression
    - SGDClassifier

Evaluation of Classifiers

- Trained with different number of features from PCA
- K-Fold Cross Validation
- Accuracy figures
  - Precision
  - Recall

Evaluation of Classifiers





Tools and Libraries used

- Numpy : Python fundamental package for scientific computing
- Pandas : Python data wrangling library
- Scikit-Learn : For Machine Learning and Vectorization
- Networkx : For graphs structure implementation
- Community : Community Detection API that works with networkx.
- Gephi : Graph Visualization Program
- Vincent : Python Plotting Library
- Jupyter Notebook : Interactive Python Programming server-client interface.

THANK YOU !!