ELEN E6893
Big Data Analytics

Research Project Proposal:
Spotify Classifier

Alex Thornton, Elmira Aliyeva, Tanvi Pande

Date Performed: November 5, 2021

Project Overview

- Create model for unsupervised music genre classification
- 10,000+ songs collected via Spotify API
- End product will accept preprocessed song metadata and predict genre
- Can k-means clusters align with genre labels?

Fig 1. Spotify - source for song metadata

Methods (Tools + Algorithms)

- Spotify Developer API
  - Source song metadata / features
  - Find songs from recommendations by genre
- Google Cloud Platform
  - Run jobs to train model with Dataproc
  - Store processed song data with BigQuery
- Spark ML Library
  - Use built in spark tools (MapReduce, k-means) for model pipeline
- GitHub
  - Shared repository for collaboration among teammates
  - https://github.com/athornton1618/SpotifyClassifier

Fig 4. Tools (GCP, GitHub, Spark)

Data V's

- **Volume**
  - 10,000+ songs will be processed
- **Variety**
  - All song metadata structured in json format
- **Velocity**
  - Data not time sensitive/ streamed
  - Spotify queried for song metadata in batches by genre


```
{
  "audio_features": [
    {
      "acousticness": 0.00242,
      "analysis_url": "https://api.spotify.com/v1/audio-analysis/2
      "danceability": 0.585,
      "duration_ms": 237040,
      "energy": 0.842,
      "id": "2takcwOaAZWiXQijPHIx7B",
      "instrumentalness": 0.00686,
      "key": 9,
      "liveness": 0.0866,
      "loudness": -5.883,
      "mode": 0,
      "speechiness": 0.0556,
      "tempo": 118.211,
      "time_signature": 4,
      "track_href": "https://api.spotify.com/v1/tracks/2takcwOaAZW
      "type": "audio_features",
      "uri": "spotify:track:2takcwOaAZWiXQijPHIx7B",
      "valence": 0.428
    }
  ]
}
```


Fig 2. Song metadata

Model Architecture

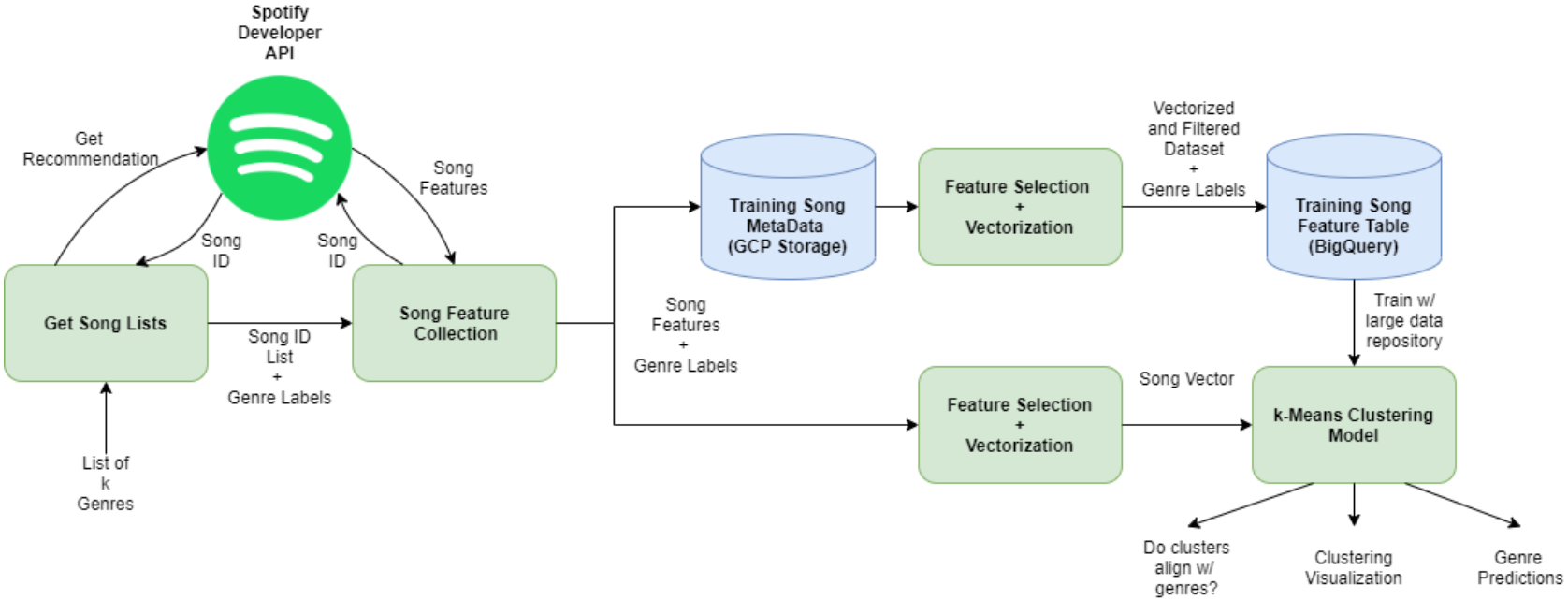

Fig 3. System Diagram

Reference Material

[1] Automatic Musical Genre Classification Of Audio Signals - Princeton CS dept

- Princeton research focused on DSP of raw audio
- Spotify won't provide audio mp3's, only preprocessed metadata with features
- Our project will scale for much larger song datasets

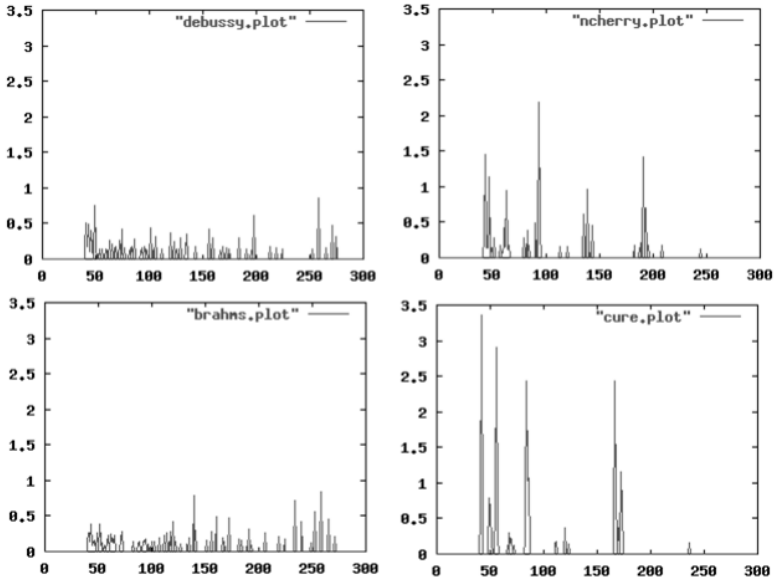

Fig 5. Beat Histogram for Classical (Left) and Pop (Right) [1]

Schedule

11/05	Proposal Presentation (Today)	Everyone
11/14	Initial Progress Checkpoint (API keys, sample song metadata json stored in GCP)	Alex
11/19	Progress Presentation	Everyone
11/28	Minimum Viable Product (product owners listed) 1. Data collection 2. Song vectorization + processing 3. K-means model 4. Analysis + Visualization	1. Alex 2. Tanvi 3. Elmira 4. Everyone
12/03	Progress Report	Everyone
12/05	Completion of all scripts, model trained, verification	Everyone
12/17	Final Slides Submission	Everyone