DJ Byg Data

Diego
Trajche
Tribhuvanesh

Introduction & Data

- Goal: Is it possible to determine the genre of a song based solely on its lyrics?

- Data sources:
  - Million Songs Dataset (Song metadata)
  - musiXmatch Dataset (Lyrics)
  - Last.fm Dataset (Tags)

- DJ Tini Data: Sample of 10,000 songs used for the development of the app.

Introduction & Data

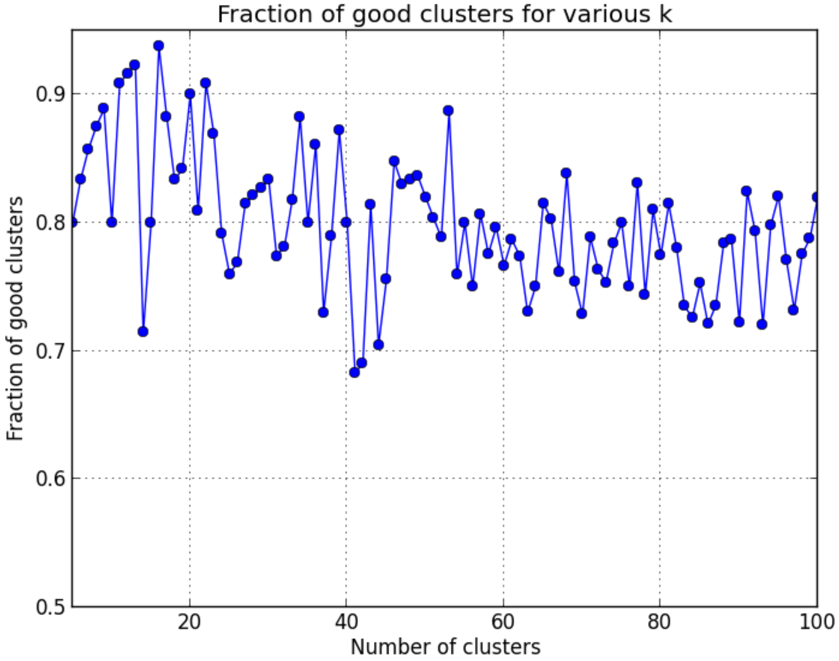
- Preprocessing
  - Convert Last.fm Tags to Genres
  - MSD ⋈ musiXmatch ⋈ Last.fm → Vector space

- Algorithms
  - K-means
  - Locality Sensitive Hashing

- Quality metrics
  - Run time
  - Accuracy - Majority Genre Probability

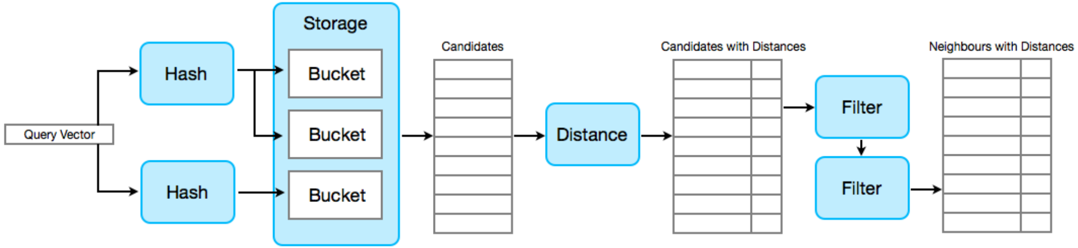
K-Means

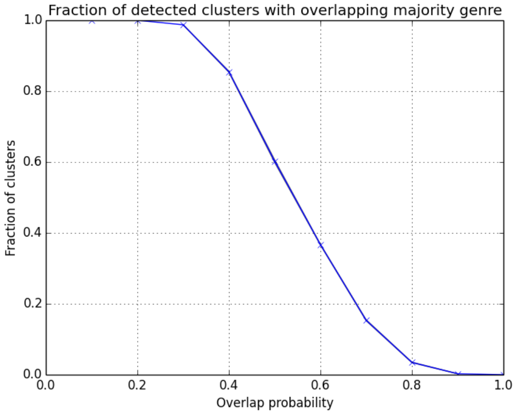
- Running time in the order of seconds
- There is a high percentage of “good clusters”
- Most of the clusters are classified as Rock
- However, it can also identify Latin, Hip-Hop and World genres



Locality Sensitive Hashing

- Similar tracks hashed to same bucket
- NearPy for LSH
- Streaming capable
- Too many parameters!
- No out-of-the-box distributed implementation
- 0.2 secs for 1.4k tracks





What next?

- 1 Million songs, 250gb dataset
- K-means (Coresets) on Hadoop
- Custom LSH on Spark
- Better quality metrics
- Does more data help?

Demo