U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

VerseVault

Your Key to Lyrics, Your Door to Music

André Lima - 202008169
Guilherme Almeida - 202006137
Jorge Sousa - 202006140
José Castro - 202006963

Introduction - **What is it?**

- **Lyric-based** Search Engine for **finding songs**
- Aggregates **various metadata** regarding musical content
- Aims to:
  - Allow **text-based querying** with the aid of context-sensitive **filters**
  - Observe a specific song’s **composition** (both **lyrically** and **structurally**)

Introduction - **Why do it?**

Lyrical and structural musical data is **vast but mostly unexplored**

Centralized and rich data sources for our data collection

In its final stage, it will be a **fun and interactive way** to interact and **perceive music production and its history**

Introduction - **What will it do?**

Basic Query Example:

“take my hand”

Advanced Query Example:

“love” **IN**: chorus **DURING**: 1970-2000 **GENRE**: disco

Tooling

mongoDB®

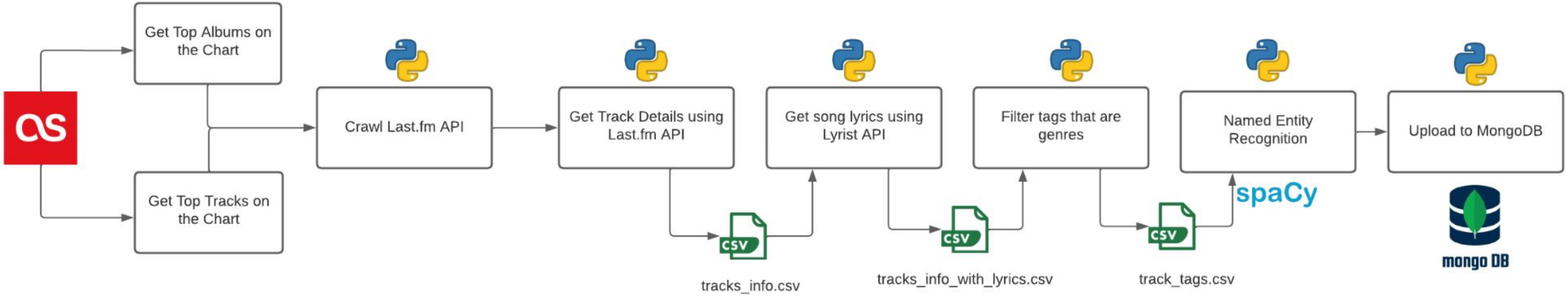
python™

matplotlib

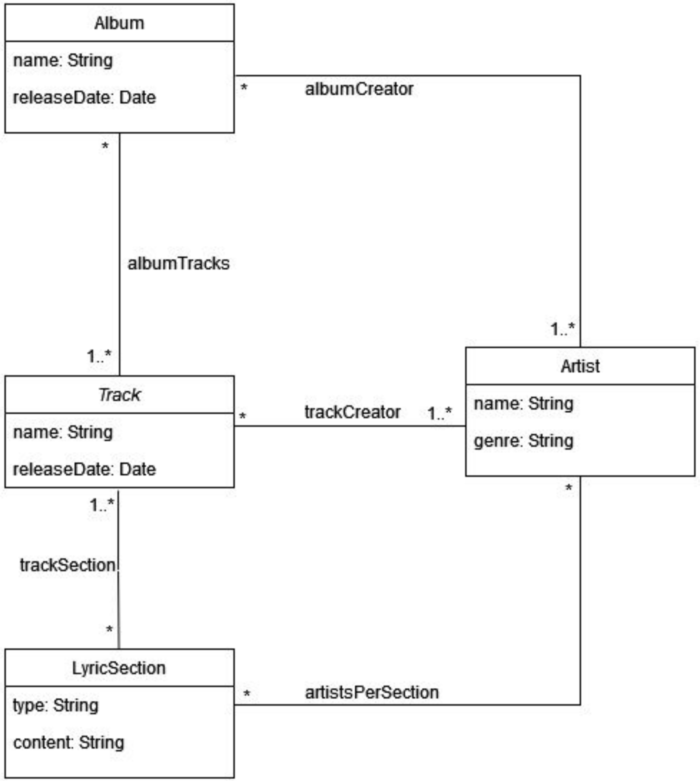
spaCy

pandas

Data Sources & Pipeline


Get Top Albums on the Chart
Get Top Tracks on the Chart
Crawl Last.fm API
Get Track Details using Last.fm API
Get song lyrics using Lyrist API
Filter tags that are genres
Named Entity Recognition
Upload to MongoDB
CSV
tracks_info.csv
CSV
tracks_info_with_lyrics.csv
CSV
track_tags.csv
spaCy
mongo DB

Domain Model



Data Collection and Processing

Data collection:

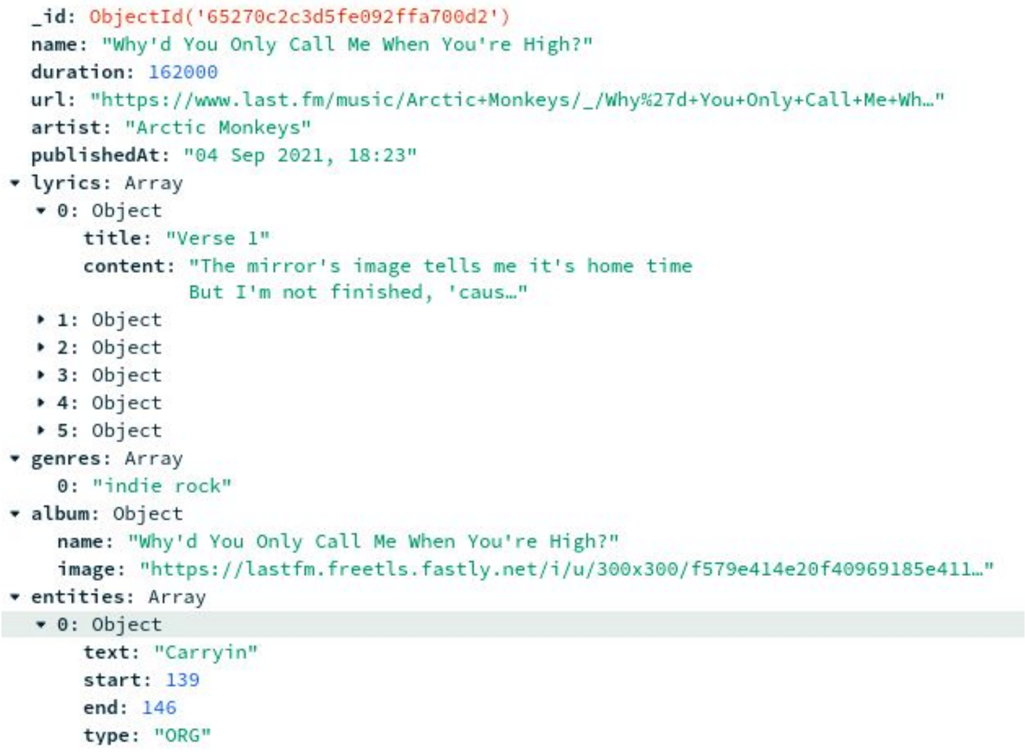
- **Last.fm API** to collect the top charting tracks and albums and their metadata, such as track genre and wiki published date. We also collect the artist and the album information of each track.
- **Lyrist API** (self-hosted) to collect the lyrics for each track obtained previously.

Data processing:

- We assign a genre to track by filtering its associated tags.
- To perform named-entity recognition in the lyrics, we use spaCy.

Finally, we upload all this information to a MongoDB instance.

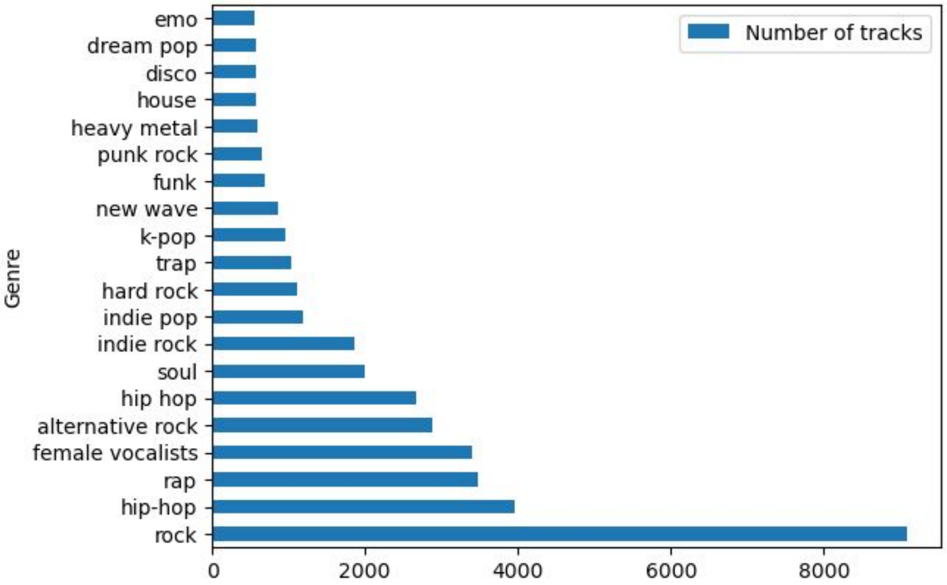
Data Collection - Example



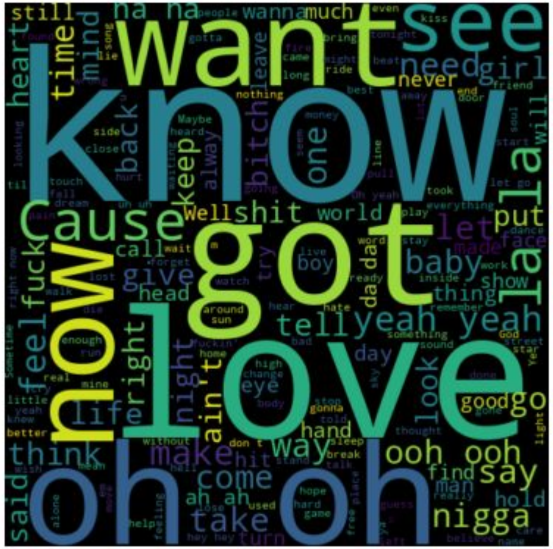
```
  _id: ObjectId('65270c2c3d5fe092ffa700d2')
  name: "Why'd You Only Call Me When You're High?"
  duration: 162000
  url: "https://www.last.fm/music/Arctic+Monkeys/_/Why%27d+You+Only+Call+Me+Wh…"
  artist: "Arctic Monkeys"
  publishedAt: "04 Sep 2021, 18:23"
▾ lyrics: Array
  ▾ 0: Object
      title: "Verse 1"
      content: "The mirror's image tells me it's home time
                But I'm not finished, 'caus…"
  ▸ 1: Object
  ▸ 2: Object
  ▸ 3: Object
  ▸ 4: Object
  ▸ 5: Object
▾ genres: Array
    0: "indie rock"
▾ album: Object
    name: "Why'd You Only Call Me When You're High?"
    image: "https://lastfm.freetls.fastly.net/i/u/300x300/f579e414e20f40969185e411…"
▾ entities: Array
  ▾ 0: Object
      text: "Carryin"
      start: 139
      end: 146
      type: "ORG"
```

Data Characterization

Genre Distribution

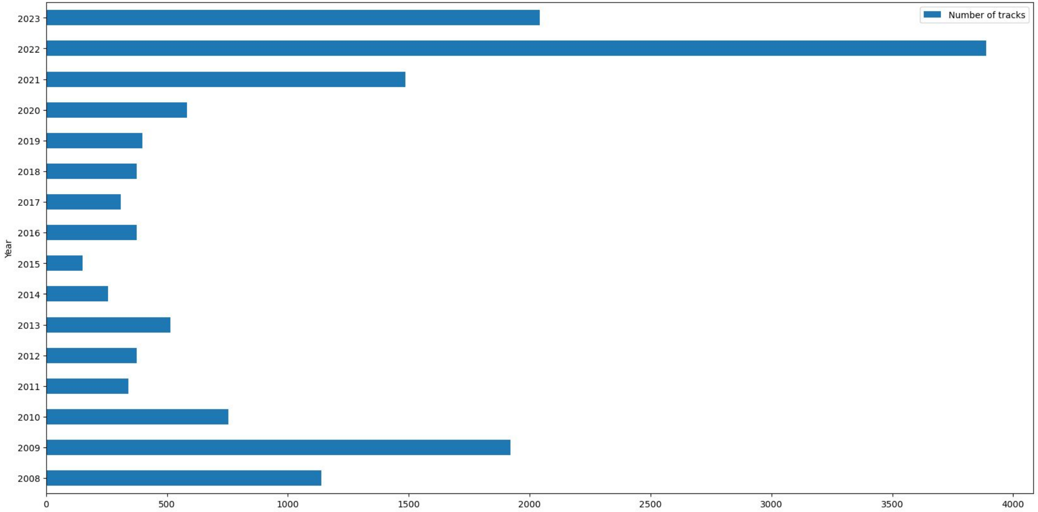


Word Cloud



Data Characterization

Number of track (per year)



Missing data (% per document)

- Missing publishing date of wiki: 68.27%
- Missing album: 8.80%
- Missing album’s image: 3.04%
- Missing sections with title: 12.79%
- Missing duration: 20.66%

Information Needs

Relevant search scenarios:

- I want to find a song containing a specific **textual excerpt**.
- I want to find a track that contains a specific **structural section**.
- I want to find a track containing a specific **textual excerpt inside a particular structural section**.
- I want to find a song **written by** a **specific artist**.
- I want to find a track **belonging to** a **specific album**.