Eksploracje Danych

Mateusz Biedak

Artur Szydek

Milionsongdataset Database

1. Cel projektu

Celem projektu była analiza bazy danych z dużą ilością piosenek, które posiadają wiele parametrów. Analiza piosenek miała polegać na znajdywaniu podobnych do siebie piosenek na podstawie tych właśnie parametrów. W pierwszym kroku należało znaleźć piosenki podobne do piosenki zadanej programowi, a w drugim podzielić je na klastry zawierające podobne do siebie piosenki tak aby mniej więcej było można wyróżnić w nich gatunki.

2. Tabele

a. Artist

- a_id (id of artist in database) [int]
- artist_name (artist name) [text]
- artist_id (artist Echo Nest ID) [text]
- artist_mbid (artist musicbrainz.org ID) [text]
- artist_playmeid (artist playme.com ID) [int]
- artist_7digitalid (artist 7digital.com ID) [int]
- artist_mbtags (tags from musicbrainz.org) [blob]
- artist_mbtags_count (tags count from musicbrainz.org) [blob]
- artist_terms (tags from Echo Nest) [blob]
- artist_terms_freq (tags frequencies from Echo Nest) [blob]
- artist_terms_weight (tags weights from Echo Nest) [blob]
- artist_familiarity (algorithmic estimation) [real]
- artist_hotttnesss (algorithmic estimation) [real]
- artist_latitude (latitude) [real]
- artist_location (location) [text]
- artist_longitude (longtitude) [real]

b. Song

- s_id (id of song in database) [int]
- artist_id (id of song's author in database) [int]

- title (song title) [text]
- year (song release year) [int]
- audio_md5 (audio hash code) [text]
- song_id (song Echo Nest ID) [text]
- release (album name) [text]
- release_7digitalid (album 7digital.com ID) [int]
- track_id (track Echo Nest ID) [text]
- track_7digitalid (track 7digital.com ID) [int]
- song_hotttnesss (algorithmic estimation) [real]
- analysis_sample_rate (sample rate of audio used) [real]
- danceability (algorithmic estimation) [real]
- duration (in seconds) [real]
- loudness (overall loudness in dB) [real]
- end_of_fade_in (seconds at the beginning of song) [real]
- start_of_fade_out (time in sec) [real]
- key (key song is in) [int]
- key_confidence (confidence measure) [real]
- energy (energy from listener point of view) [real]
- mode (major or minor) [int]
- mode confidence (confidence measure) [real]
- tempo (in BPM) [real]
- time_signature (estimate of number of beats per bar) [int]
- time_signature_confidence (confidence measure) [real]
- sections_start (largest grouping in a song) [blob]
- sections_confidence (confidence measure) [blob]
- bars_start (beginning of bars) [blob]
- bars_confidence (confidence measure) [blob]
- beats_start (result of beat tracking) [blob]
- beats_confidence (confidence measure) [blob]
- segments_start (musical events) [blob]
- segments_confidence (confidence measure) [blob]
- segments_loudness_max (max dB value) [blob]
- segments_loudness_max_time (time of max dB value) [blob]
- segments_loudness_max_start (dB value at onset) [blob]
- segments_piches (chroma feature) [blob]
- segments_timbre (texture features) [blob]
- tatums_start (smallest rythmic element) [blob]
- tatums_confidence (confidence measure) [blob]

c. ArtistsRel

- id (id of row in database) [int]
- artist1_id (id of selected artist in database) [text]
- artist2_id (id of similar artist in Echo Nest) [text]

3. Statystyki

Zgromadziliśmy trzy główne statystyki na tej bazie danych. A oto ich przykłady:

- Ile piosenek ma każdy artysta:

David Taylor - 1 song(s)

David Zinman;Pittsburgh Symphony Orchestra - 2 song(s)

Davol - 2 song(s)

Dax Riggs - 4 song(s)

Daylight Torn - 1 song(s)

De La Ghetto - 1 song(s)

De La Soul / MF Doom - 2 song(s)

DeGarmo & Key - 5 song(s)

Dead Hearts - 1 song(s)

Dead Kennedys - 3 song(s)

Dead Prez - 2 song(s)

Dead To Me - 2 song(s)

Dealership - 1 song(s)

Dean Elliott And His Big Band - 1 song(s)

Dean Evenson - 1 song(s)

Dean Martin - 3 song(s)

Death Cab for Cutie - 1 song(s)

Death From Above 1979 - 2 song(s)

- Ile albumów ma każdy artysta:

Bowen_ Robin Huw - 1 album(s)

Boyz II Men - 3 album(s)

Brainchoke - 1 album(s)

Brand X - 2 album(s)

Bravehearts - 1 album(s)

Brazilian Tropical Orchestra - 5 album(s)

Brenda Boykin - 1 album(s)

Brenda Lee - 5 album(s)

Brent Lamb - 2 album(s)

Brian Auger's Oblivion Express - 1 album(s)

Brian Dullaghan - 3 album(s)

Brian Eno And David Byrne - 1 album(s)

Brian Free & Assurance - 2 album(s)

Brian Keane - 2 album(s)

Brian Littrell - 1 album(s)

Brian Tyler - 2 album(s)

Brigada Victor Jara - 1 album(s)

Britney Spears - 9 album(s)

Britny Fox - 1 album(s)

Britt Nicole - 4 album(s)

Brixx - 1 album(s)

- Ile piosenek ma każdy album:

The Paul - EP - 1 song(s)

The Paul Butterfield Blues Band Live - 1 song(s)

The People Vs. - 2 song(s)

The People or The Gun - 2 song(s)

The Perfect Blend - 1 song(s)

The Pink & The Lily - 1 song(s)

The Piper At The Gates Of Dawn - 1 song(s)

The Plan - 5 song(s)

The Planxty Collection - 1 song(s)

The Platinum Collection - 7 song(s)

The Poems of Elizabeth Bishop and Other Songs - 1 song(s)

The Poison - 2 song(s)

The Polar Express - Original Motion Picture Soundtrack - 1 song(s)

The Preacher's Son - 1 song(s)

4. Porównywanie piosenek

Aby porównywać piosenki, musieliśmy stworzyć swoją własną metrykę, która liczyła różnice między wartościami tych parametrów z bazy danych, które nie są tablicowymi, ponieważ ich porównywanie nie jest możliwe bez odpowiedniej wiedzy technicznej na temat muzyki. Natomiast reszta parametrów, została zsumowana ze sobą z odpowiednimi wagami wyznaczonymi doświadczalnie oraz przy użyciu rozsądku i logiki. Uprzednio jednak zostały one poddane normalizacji aby liczby, które są duże tak jak długość trwania piosenki wyrażona w sekundach, nie wpływały tak bardzo na wyniki.

Parametry, które uwzględnia Nasza metryka to (wraz z wagami):

- title (0.02)
- year (0.55)
- song_hotttnesss (0.08)
- danceability (0.77)
- duration (0.04)
- loudness (0.38)
- end_of_fade_in (0.01)
- start_of_fade_out (0.03)
- key (0.05)
- energy (0.1)
- mode (0.2)
- tempo (0.6)
- time_signature (0.3)
- artist_familiarity (0.9)
- artist_hotttnesss (0.96)
- artist_latitude (0.5)
- artist_location (0.01)
- artist_longitude (0.3)

Z racji tego, że im metryka mniejsza tym piosenki są bardziej do siebie podobne, niektóre wartości dostawały duże wagi, aby po danej cesze łatwiej je było rozróżnić i ich różnica była bardziej wyraźna, a gdy cecha ma mały wpływ na różnicę, dostawała małą wagę, aby nie zmieniała ona za dużo, a tylko kosmetycznie.

Oprócz tych parametrów, sprawdzane było także to, czy dane dwie piosenki znajdują się na tym samym albumie – jesli tak to całą wartość metryki jest mnożona przez 0.03 bo napewno są do siebie tematycznie podobne, gdy piosenki były tego samego artysty to mnożnik metryki wynosił 0.04, ponieważ między albumami mogą istnieć pewne różnice ale całość twórczości artysty brzmi podobnie, natomiast gdy artysta jednej pisoenki był na liście artystów podobnych do artysty drugiej piosenki, to mnożnik wynosił 0.21, ponieważ nie powinno to tak bardzo wpływać na

podobieństwa piosenek, ale pewne podobieństwo gatunkowe zostało zauważone więc powinno być uwzględnione.

Na samym końcu porównania Naszej piosenki z wszystkimi w bazie, wypisywane jest na ekran 20 najbliższych piosenek, zazwyczaj najbliższymi są te, które należą do tego samego artysty.

Przykład dla piosenki „The Deceived" artysty „Trivium":

Similar songs to The Deceived:

Upon The Shores [Explicit] : 0.359087952842
Into The Mouth Of Hell We March [Explicit] : 1.02065124741
Shogun [with fade_ for special edition] : 1.92227848439
Left For Dead (Album Version) : 1.98802889926
Machine Gun Majesty [Live] (Album Version) : 2.8512859075
Deviate From The Form : 3.06450021763
Destroy Everything (Album Version) : 3.09556197587
The End Of The Line : 3.32597617058
What You Deserve (Album Version) : 3.34492292365
The Three-Dimensional Shadow : 3.39567549181
As The Sleeper Awakes : 3.46313889679
Abstracted : 3.85426739036
All I Ask For (Album Version) : 4.27974976412
What Drives The Weak : 4.36703126273
One the road (to Damnation) : 4.40915014987
Shadows That Move : 4.64840753644
Origins And Endings : 4.94068676776
Sensory Deprivation Adventure : 5.25727570416
Into Battle : 5.3308877274
IV : 5.40413725782

5. Klasteryzacja piosenek

Klasteryzacja piosenek odbyła się najprostszym algorytmem zachłannym. Wybierana została pierwsza piosenka dla danego klastra, wszystkie inne zostawały z nią porównywane i jeśli którejś z nich metryka wyszła mniejsza niż 5.0, to była to podobna piosenka, z racji czego była ona wrzucana do tego samego klastra a usuwana ze zbioru piosenek. Działo się tak dopóki zbiór piosenek nie był pusty. Jest to dość niska metryka więc klastrów stworzyło się naprawdę dużo, najczęściej z samym artystą, choć zdarzały się większe klastry z kilkoma artystami.

Przykładowe klastry:

Killed By Love, Pursuit Of Happiness
She's So Young, Pursuit Of Happiness
Hard To Laugh, Pursuit Of Happiness
Standing Knee Deep In a River (Dying Of Thirst), Kathy Mattea
Down On Him, Pursuit Of Happiness
Elle Voulait Jouer Cabaret, Patricia Kaas
One Night Alone, Vixen

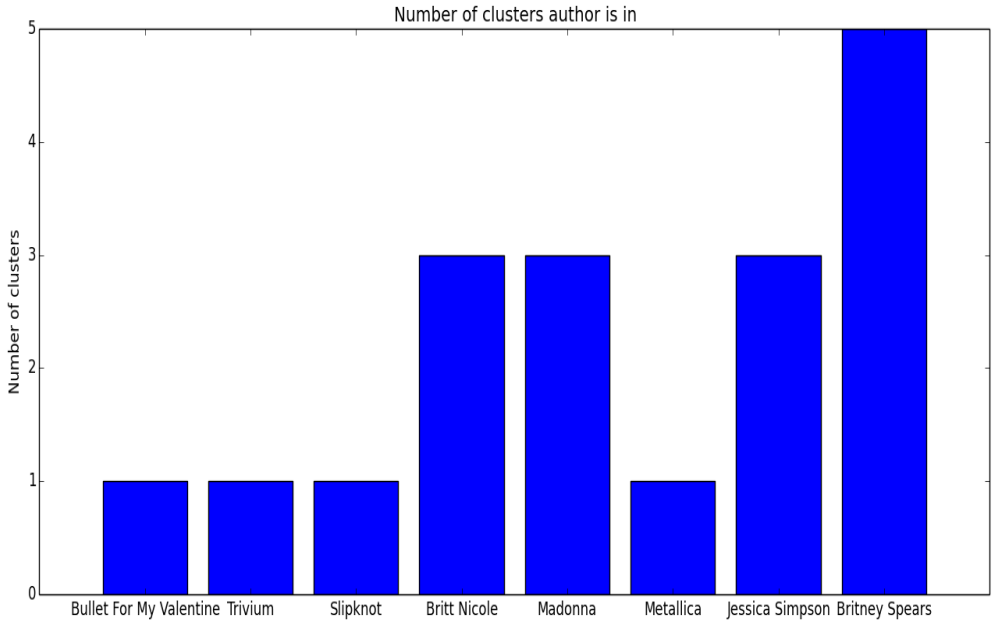
Good Girl Gone Bad, Rihanna
We Ride, Rihanna
Relax, 2-4 Grooves feat. Reki D.
SOS, Rihanna
Wake The Sleeper, Uriah Heep
Music Of The Sun, Rihanna
Loba, Shakira
Don't Stop The Music, Rihanna
Good Girl Gone Bad, Rihanna

I Know Somethin (Bout You), Alice In Chains
Brother, Alice In Chains
God Smack, Alice In Chains
Sickman, Alice In Chains
Naked In The Rain (Album Version), Red Hot Chili Peppers
Take Her Out, Alice In Chains

6. Analiza wyników

Napisany przez Nas algorytmy wyznaczające klastry oraz metryka, która porównywała bliskość piosenek wydają się być w miarę rozsądnymi ponieważ dają dość dobre wyniki. Piosenki przydzielone do klastrów są rzeczywiście piosenkami z tego samego bądź podobnego gatunku a najczęściej są to piosenki tego samego autora, choć

zdarzają się też przypadki, że jeden autor znajduje się w kilku klastrach tak jak obrazuje to poniższy wykres.



Jak widać zespoły, które są z gatunku „Rock” albo „Heavy Metal” zazwyczaj należą do jednego klastra natomiast te z gatunku „Pop” rozłożone są na kilka klastrów. Można na tej podstawie wysunąć następujące wnioski:

- artyści rockowi i metalowi są na tyle specyficzni, że są dość łatwo rozróżnialni podczas gdy popowi są na tyle do siebie podobni, że bywają myleni z innymi artystami

- artyści popowi częściej mają tendencję do „podkradania” melodii innych artystów niż artyści innych gatunków