Eksploracje Danych

Mateusz Biedak

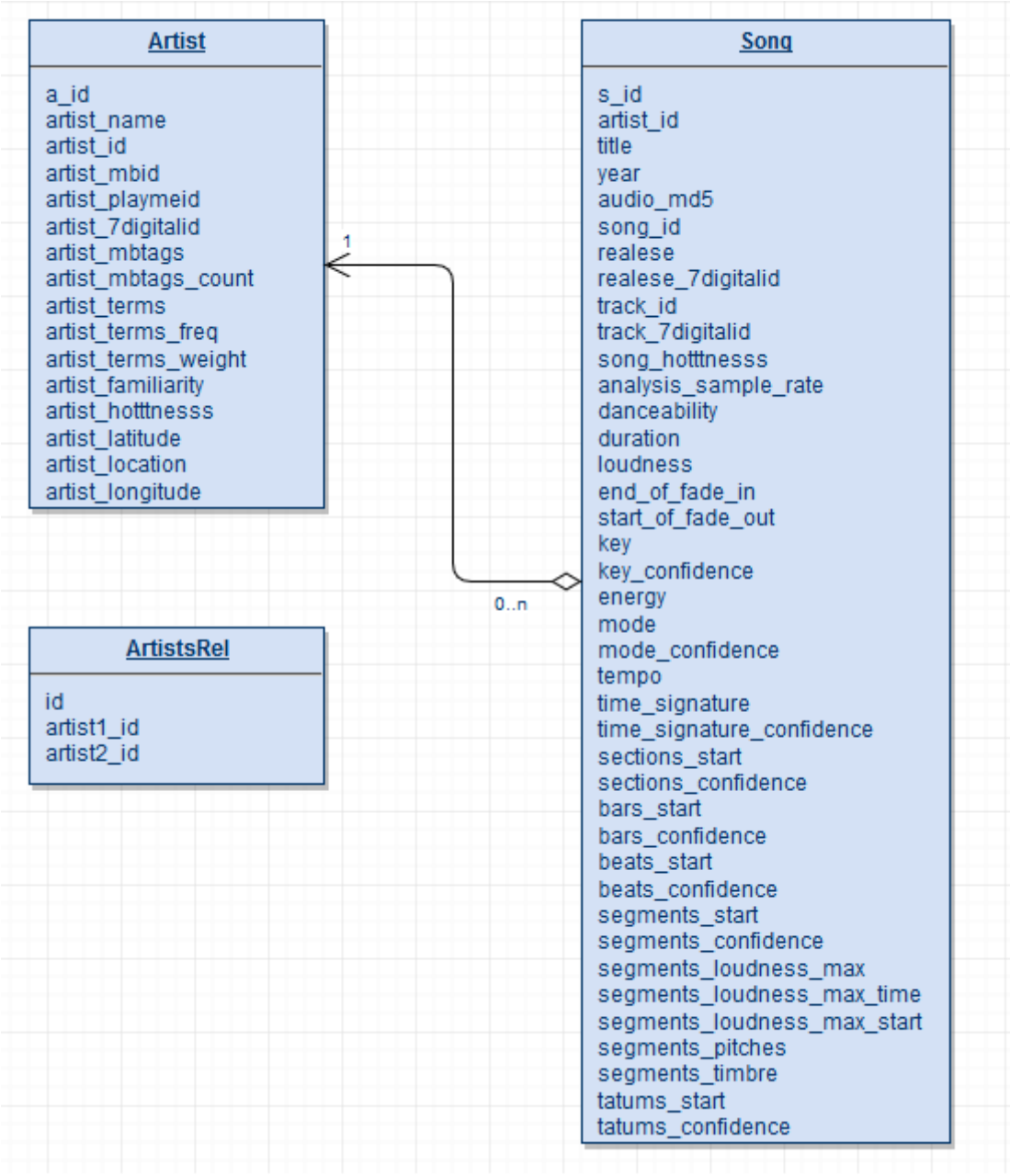
Artur Szydek

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 [int] - id artysty w bazie danych
- artist_name [text] – nazwa artysty
- artist_id [text] - id artysty w Echo Nest
- artist_mbid [text] - id artysty w musicbrainz.org
- artist_playmeid [int] – id artysty w playme.com
- artist_7digitalid [int] – id artysty w 7digital.com
- artist_mbtags [blob] – tagi artysty z musicbrainz.org
- artist_mbtags_count [blob] – liczba tagów artysty z musicbrainz.org
- artist_terms [blob] – tagi artysty z Echo Nest
- artist_terms_freq [blob] – częstość występowania tagów artysty z Echo Nest

- artist_terms_weight [blob] – waga tagów artysty z Echo Nest
- artist_familiarity [real] – algorytmiczna estymacja popularności artysty
- artist_hotttnesss [real] - algorytmiczna estymacja rozchwytywalności artysty
- artist_latitude [real] – szerokość geograficzna pochodzenia artysty
- artist_location [text] – pochodzenie artysty
- artist_longitude [real] - długość geograficzna pochodzenia artysty

b. Song

- s_id [int] - id piosenki w bazie danych
- artist_id [int] – id autora piosenki w bazie danych
- title [text] - tytuł
- year [int] – rok wydania
- audio_md5 [text] – zakodowana postać piosenki
- song_id [text] - id piosenki w Echo Nest
- release [text] – nazwa albumu
- release_7digitalid [int] – id albumu z 7digital.com
- track_id [text] – id ścieżki z Echo Nest
- track_7digitalid [int] – id ścieżki z 7digital.com
- song_hotttnesss [real] – algorytmiczna estymacja rozchwytywalności utworu
- analysis_sample_rate [real] – ilość próbek piosenki użytych do analizy
- danceability [real] – algorytmiczna esymacja taneczności utworu
- duration [real] – czas trwania
- loudness [real] – głośność w decybelach
- end_of_fade_in [real] – czas wstępu do piosenki
- start_of_fade_out [real] – czas zakończenia piosenki
- key [int] – klucz piosenki
- key_confidence [real] – oszacowana pewność klucza
- energy [real] – energia z punktu widzenia słuchacz
- mode [int] – skala w muzyce – major albo minor
- mode confidence [real] – oszacowana pewność skali
- tempo [real] – tempo piosenki w BPM
- time_signature [int] – estymata ilości uderzeń danego chwytu
- time_signature_confidence [real] – oszacowanie pewności ilości uderzeń
- sections_start [blob] – sekcje w piosence
- sections_confidence [blob] – oszacowanie pewności podziału na sekcje
- bars_start [blob] – momenty rozpoczęcia chwytów
- bars_confidence [blob] – oszacowanie pewności rozpoczęcia chwytów
- beats_start [blob] – momenty rozpoczęcia uderzeń
- beats_confidence [blob] – oszacowanie pewności rozpoczęcia uderzeń
- segments_start [blob] – części z których składa się piosenka
- segments_confidence [blob] – oszacowanie pewności części piosenki
- segments_loudness_max [blob] – maksymalna głośność każdej części
- segments_loudness_max_time [blob] – czas trwania maksymalnej głośności

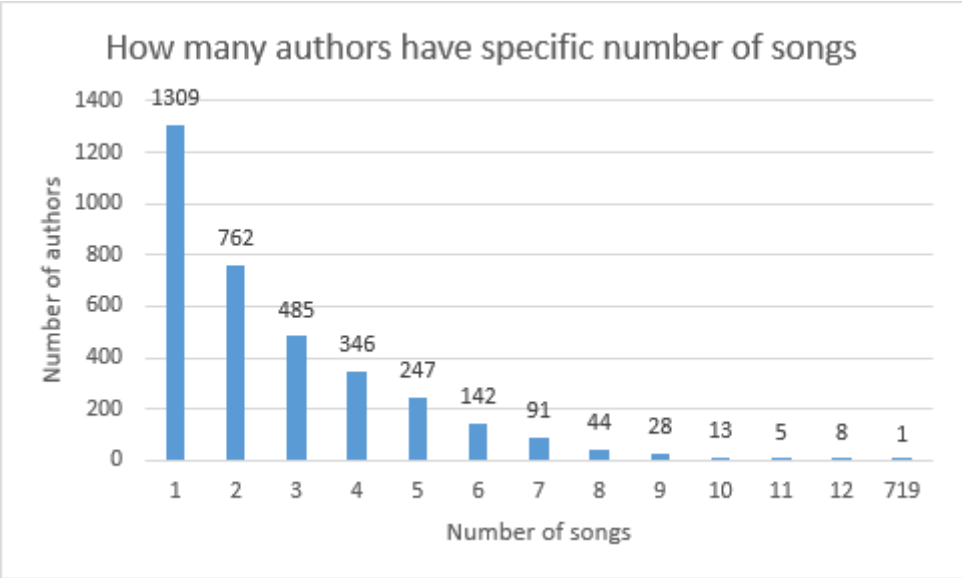
- segments_loudness_max_start [blob] – rozpoczęcie maksymalnej głośności
- segments_piches [blob] – wysokości tonów w częściach piosenki
- segments_timbre [blob] – barwy dźwięku w częściach piosenki
- tatums_start [blob] – najmniejsze rytmiczne elementy piosenki
- tatums_confidence [blob] – oszacowanie pewności najmniejszych elementów piosenki

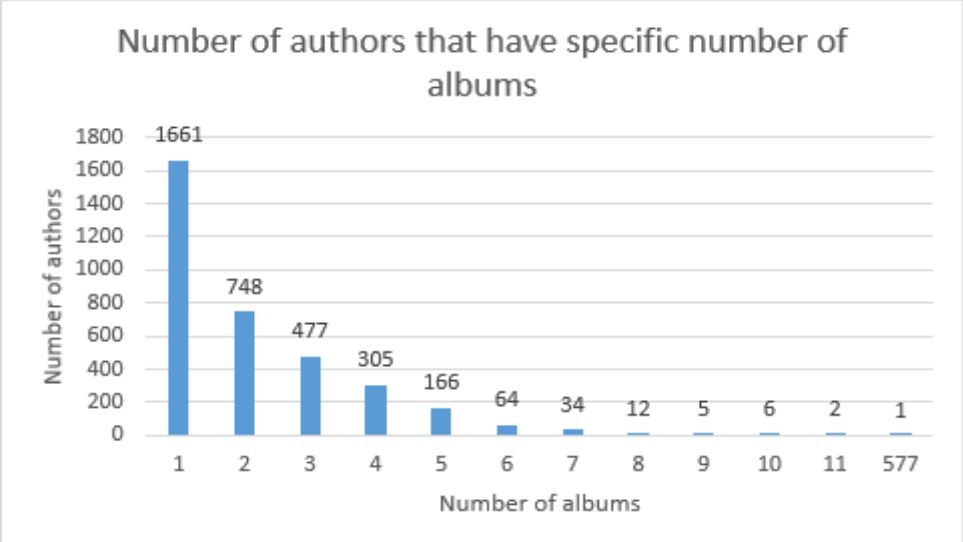
c. ArtistsRel

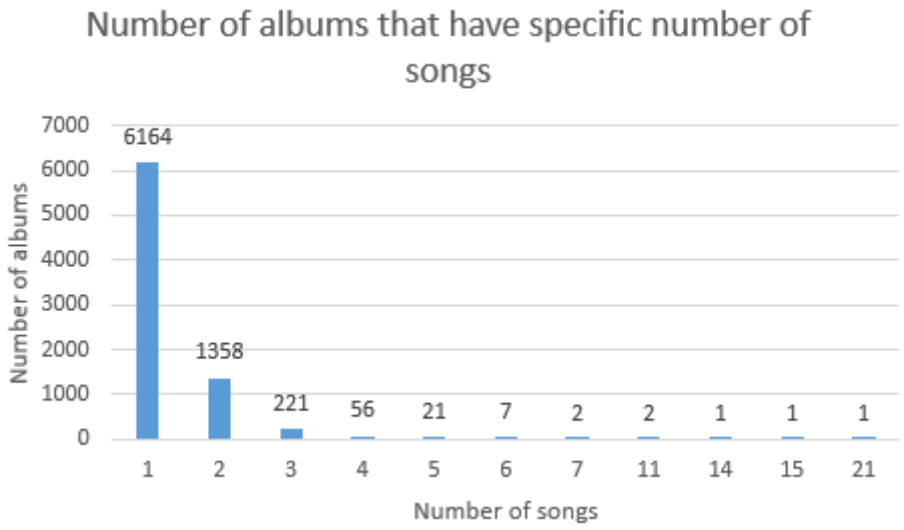
- id [int] – id rekordu w bazie danych
- artist1_id [text] – id wybranego artysty w bazie danych
- artist2_id [text] – id artysty z Echo Nest podobnego do artysty 1

3. Statystyki

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







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 biorąc pod uwagę to jaka wartość może mieć rzeczywiście wpływ na to, że piosenki są do siebie podobne. Na przykład tytuł nie będzie miał większego znaczenia więc dostał on małą wagę, a klucz oraz tempo czy taneczność piosenki już jakiś wpływ może mieć.

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 (0.5)
- 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)

Metryka działa więc tak, że wszystkie wyżej wymienione parametry dla obu porównywanych piosenek są normalizowane do przedziału [0, 1], następnie liczona jest różnica tych samych parametrów. Każda z tych różnic jest mnożona przez odpowiednią wagę, a następnie wszystkie różnice są ze sobą sumowane.

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

5. Klasteryzacja piosenek

Klasteryzacja piosenek odbyła się najprostszym algorytmem zachłannym. Tworzony był pusty klaster a do niego wrzucana pierwsza z brzegu piosenka. Wszystkie inne zostawały z nią porównywane i jeśli którejś z nich metryka wyszła mniejsza niż pewien parametr odcięcia, 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. Biorąc za parametr odcięcia wartość 5.0 klastrów stworzyło się naprawdę dużo, najczęściej z samym artystą, choć zdarzały się większe klastry z kilkoma artystami.

Przykładowe klastry:

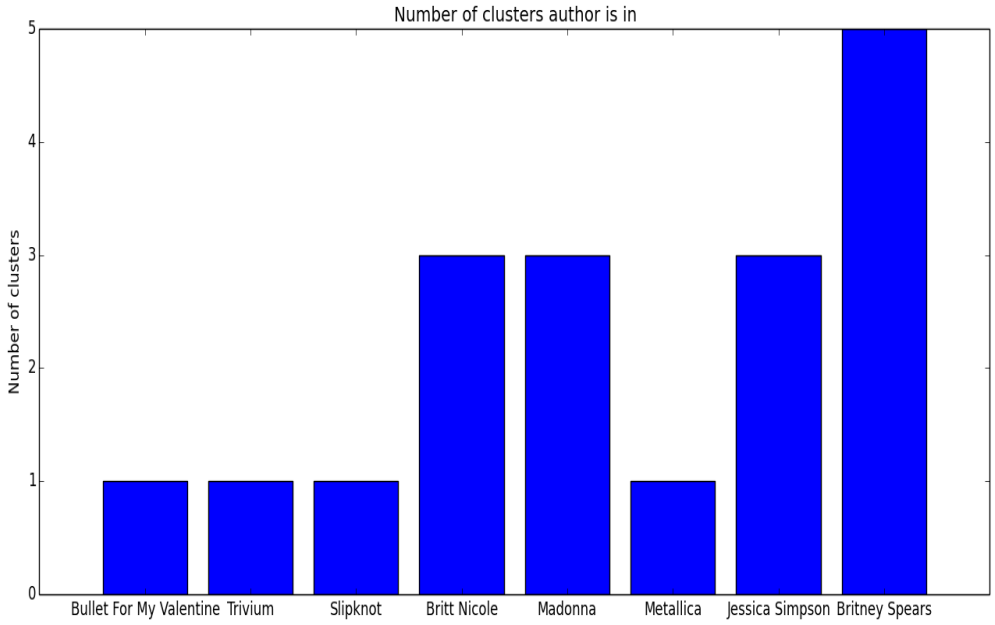
Killed By Love, Pursuit Of Happiness
She's So Young, Pursuit Of Happiness
Hard To Laugh, Pursuit Of Happiness
Standing Knee Deep In a River (Dying Of Thirst), Kathy Mattea
Down On Him, Pursuit Of Happiness
Elle Voulait Jouer Cabaret, Patricia Kaas
One Night Alone, Vixen

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

7. Podsumowanie projektu

Projekt był średnio ciekawy, może z tego względu, że nie jesteśmy ekspertami w dziedzinie muzyki i nie potrafiliśmy użyć wszystkich danych, które prezentowała baza, aby sprawić by to co było liczone przez program mogło być przez kogoś rzeczywiście używane, jesto to raczej narzędzie, którym można na oko sprawdzić jacy artyści są do siebie podobni i ewentualnie, które ich piosenki są warte przesłuchania jeśli podoba nam się jedna z piosenek danego autora.

Nie jest to niestety narzędzie, którego można by było użyć do wykrywania plagiatów w piosenkach które nimi rzeczywiście są, ale jakaś drobna rzecz została w nich zmieniona, czy to tempo czy instrumenty na których piosenka jest grana, czy też sam gatunek.

8. Możliwe rozszerzenia w przyszłości

Możliwe rozszerzenia to:
- zdobycie wiedzy na temat muzyki i podłączenie pod metrykę danych tablicowych, których my nie potrafiliśmy zrozumieć
- ulepszenie porównywania piosenek o to, żeby nie były porównywane tylko jako całość, ale też jako konkretne sekcje, aby wykryć częściowy plagiat
- dopisanie interfejsu użytkownika, który mógłby po wpisaniu danej piosenki, podać piosenki podobne – coś podobnego do spotify, ponieważ na razie wszystko jest wypisywane na konsolę.