Przeszukiwanie zespołów muzycznych w poszukiwaniu ciekawych zależności

Mateusz Kruszyna `inf127252` Bartosz Górka `inf127228`
Jarosław Skrzypczak `inf127265`

17 czerwca 2018

1 Wprowadzenie

W realizowanym eksperymencie podjęliśmy się analizy zależności, jakie można odkryć w danych dotyczących zespołów muzycznych. Postanowiliśmy ograniczyć się do dwóch kategorii, aby przedstawić wybrane charakterystyki, bez nadmiarowych analiz.

2 Pozyskanie danych

Do pozyskania danych do analizy zastosowano *Apache Solr* oraz *Apache Nutch*. Postanowiono wykorzystać strony *Wikipedii* i czternaście z nich wstawić do zbioru startowych odnośników. Koniecznym było zagwarantowanie poprawnych ustawień w silnikach przetwarzających, aby ignorowały one niepotrzebne odnośniki takie jak zewnętrzne strony, załączniki, historie edycji bądź też odnośniki do mediawiki.

Po przygotowaniu środowiska, udało się pozyskać informacje w postaci 1050 dokumentów na temat zespołów muzycznych. Niestety, po wstępnej analizie zebranych informacji okazało się, że część odnośników prowadziła do podkategorii, których zespoły zostały uwzględnione w bazie. Postanowiono wyeliminować strony podkategorii.

Tak uzyskany indeks (zbudowanego z wykorzystaniem *Apache Solr* + *Apache Nutch*) poddano analizie z wykorzystaniem *Apache Lucene*. Z każdego dokumentu pobrano następujące informacje:

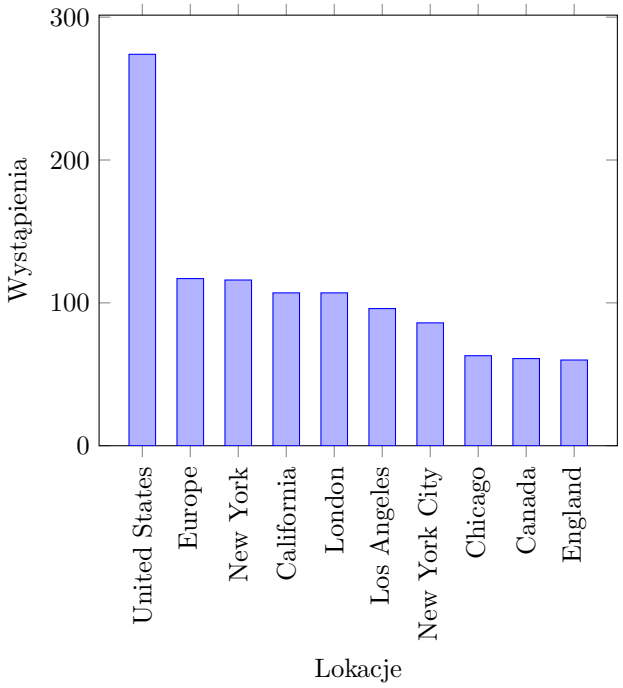
- Tytuł, wycięty z adresu URL strony
- Zawartość dokumentu (plik html) został poddany tokenizacji
- Lokalizacje - uzyskane za pomocą modułu do rozpoznawania lokalizacji *OpenNLP*.
- Kategorie zostały wybrane z przeszukiwanej bazy, dzięki charakterystycznym schematom linków, które zawierały słowo kluczowe ‘category:‘
- Pozyskiwanie dat (rok) z tekstu niestety nie okazało się poprawne przy wykorzystaniu gotowego modelu z pakietu *OpenNLP*. Jako alternatywę wykorzystano wyrażenia regularne do wykrycia dwóch zapisów zapisów lat (format XXXX oraz ’XX dla lat 19XX). Ponadto ograniczono lata do zakresu [1000, 2051]
- Wyszukiwanie imion i nazwisk zostało zrealizowane dzięki wykorzystaniu *OpenNLP*. Także tym razem wyniki nie były jednoznaczne i zawierały wiele niewłaściwych form. Z tego powodu wszelkie imiona i nazwiska, które zawierały jakieś słowo z listy (”Tools What”, ”Permanent”, ”Page”, ”The”, ”Music”, ”In”, ”Retrieved”, ”American”) nie zostały uwzględnione w zliczaniu.
- Wszelkie słowa, które były filtrem w jakimkolwiek zliczaniu czy wybieraniu danych, zostały odkryte z teksu poprzez pełne przeszukiwanie i zaobserwowanie danych anomalii.

3 Wstępne oczyszczenie danych

W lokalizacjach pominięto ‘Random‘ którą uznano jako błędnie rozpoznawaną. Gdy w nazwie kategorii pojawiło się jedno ze słów zakazanych (”by”, ”genre”, ”navigational”, ”(genre)”, ”musicians”, ”nationality”, ”body”), to taka nazwa nie była brana pod uwagę. Po takim zabiegu, aby wyszukać do jakiej

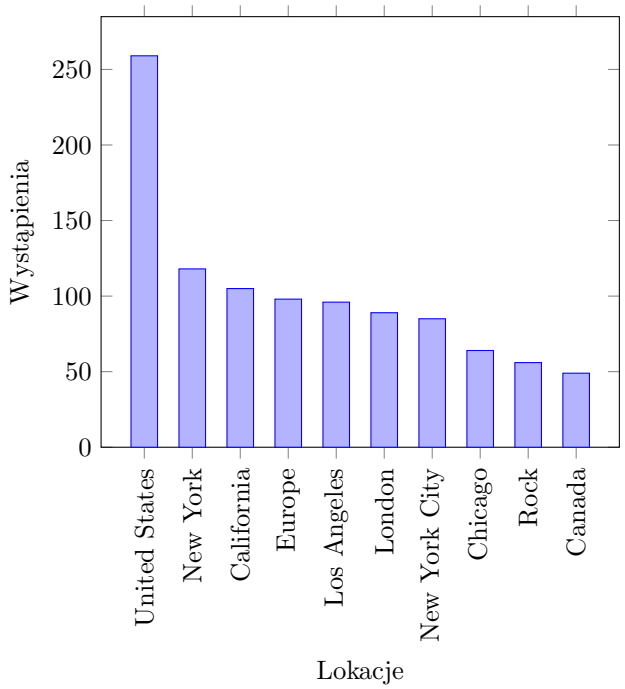
kategorii należy dokument, dla każdego dokumentu słowa kluczowe w postaci pojedynczych wyrazów z kategorii zostały zliczone sumaryczne wystąpienia każdego z tych słów w dokumencie.

4 Analiza wykresów



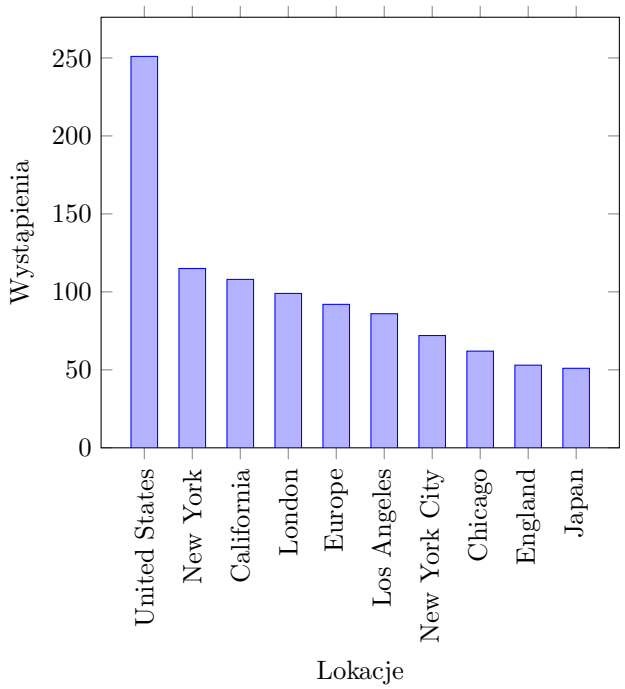
Rysunek 1: Najcześciej występujące lokalizacje dla grupy 'American blues rock'

Zgodnie z oczekiwaniami, dla zespołu odnoszącego się do Stanów Zjednoczonych, taka fraza (United States) pojawiała się najczęściej. Co jest zaskakującego - bardzo wiele wystąpień odnotowała Europa. Również i ostatnia w prezentowanym zestawieniu Wielka Brytania (England) cieszyła się popularnością mimo braku powiązania wprost. Być może fraza 'American blues rock' odniosła się także do koncertów w Europie i ich powiązań z europejskimi wykonawcami. New York (pozycja 3), California (pozycja 4), Los Angeles (pozycja 6), New York City (pozycja 7) to przykłady lokalizacji których spodziewaliśmy się osiągnąć w analizie lokalizacji.



Rysunek 2: Najczęściej występujące lokalizacje dla grupy 'American blues'

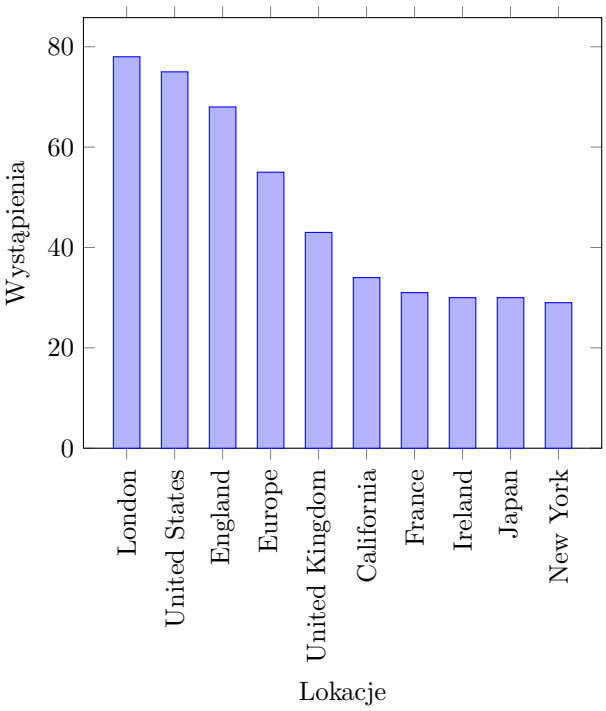
Podobnie jak w przypadku analizy dla American blues rock, również i American blues w znaczącej większości odnoszą się do Stanów Zjednoczonych. W tym przypadku mamy zmianę ustawienia lokalizacji w stosunku do wcześniej charakteryzowanego Rysunek 1. Co ciekawe, w zestawieniu pojawiła się fraza 'Rock' która jest dość sporna - teoretycznie może odnosić się zarówno do lokalizacji (góra / skała jak i stylu jakim jest rock).



Rysunek 3: Najczęściej występujące lokalizacje dla grupy 'American instrumental'

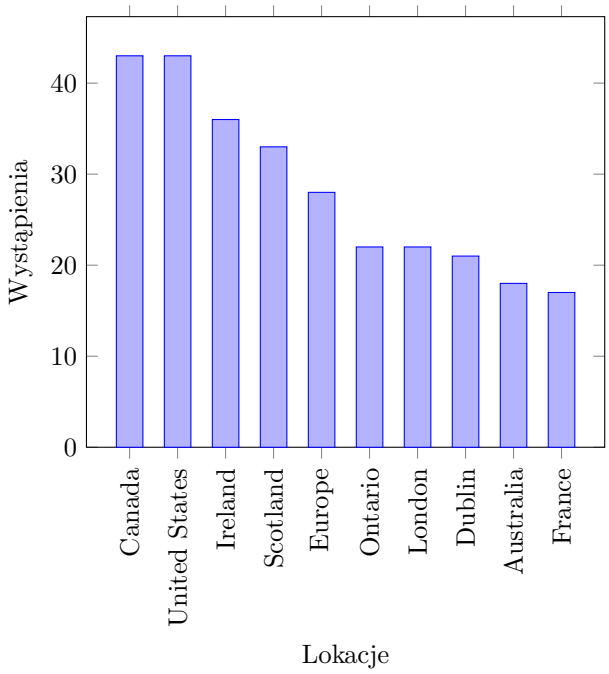
W przypadku analizy lokalizacji jakie zostały wyszukane dla 'American instrumental' poza oczywisty-

mi trendami zaprezentowanymi również w przypadku Rysunek 1 oraz Rysunek 2, mamy niespodziewane wystąpienie ‘Japan‘ czyli odniesienia do Japonii. Ciężko scharakteryzować przyczynę tego wystąpienia, może to być nawiązanie do koncertów w tamtym kraju, bądź także i wszelkie wzorce czerpane z tego kraju.



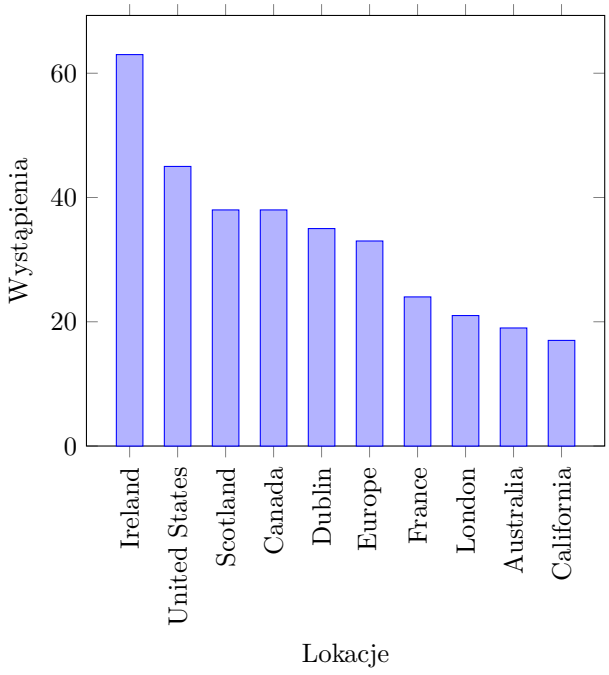
Rysunek 4: Najczęściej występujące lokalizacje dla grupy ’British instrumental’

Ciekawe wzorce odnoszące się do lokalizacji pojawiają się dla ‘British instrumental‘. Tutaj przeważa wystąpienie Londynu oraz odnośniki do europejskich nazw tj. Wielka Brytania, Europa, Zjednoczone Królestwo, Francja czy Irlandia. Na uwagę zasługują jednak odnośniki do Kalifornii, Stanów Zjednoczonych oraz Nowego Jorku - które mogą wynikać z koncertowania na terenie USA przez zespoły. Jednakże najbardziej zaskakującym jest ponownie odnośnik do Japonii. Podobnie jak w przypadku analizy dla Rysunek 3. Tutaj również ciężko scharakteryzować dlaczego tak licznie wystąpił odnośnik do kraju kwitnącej wiśni.



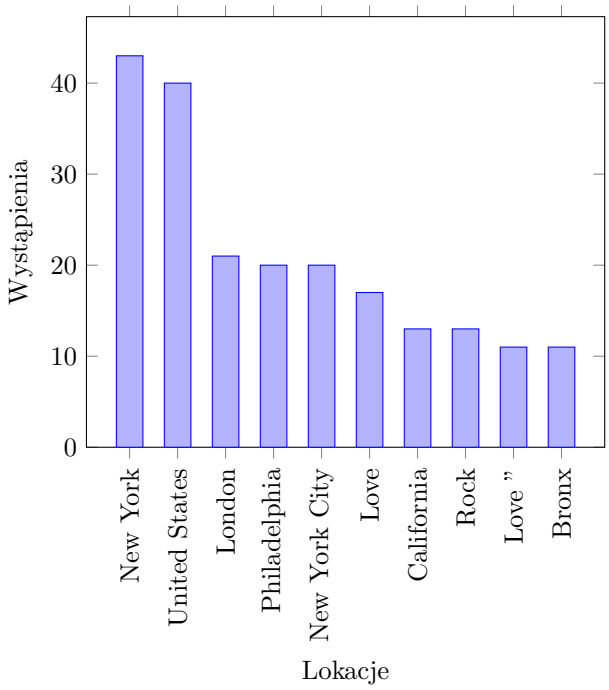
Rysunek 5: Najcześciej występujące lokalizacje dla grupy 'Canadian Celtic'

Analizując 'Canadian Celtic' możemy zauważyć jak wielkie znaczenie ma drugie określenie we frazie czyli nawiązanie do Celtów. W tym przypadku zgodnie z oczekiwaniami, na dwóch pierwszych miejscach mamy Kanadę oraz Stany Zjednoczone. Jednakże kolejne dwie pozycji (oraz dodatkowo Dublin) to ewidentny przykład nawiązania do krajów związanych z kulturą celtycką czyli Irlandii oraz Szkocji.



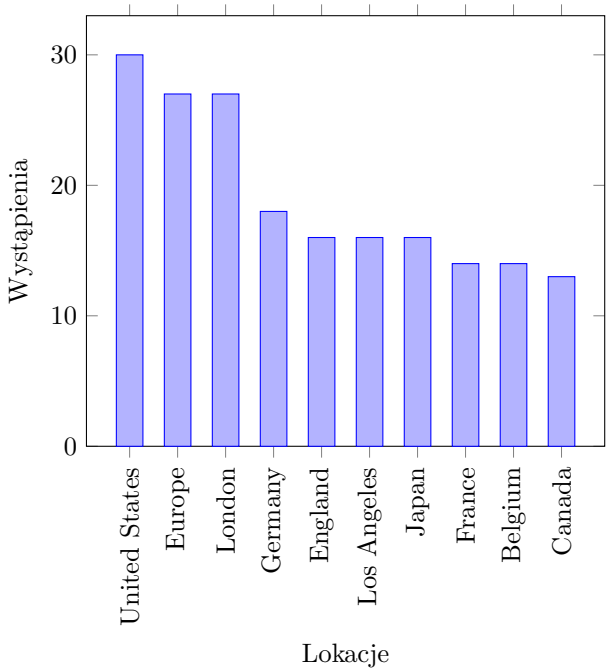
Rysunek 6: Najcześciej występujące lokalizacje dla grupy 'Celtic'

Dla samej frazy 'Celtic' mamy bardzo liczne nawiązania do Irlandii oraz Szkocji. Ponadto pojawiają się europejskie lokalizacje. Dużym zaskoczeniem niewątpliwie jest Australia, która raczej nie kojarzy się w ogóle z omawianym określeniem.



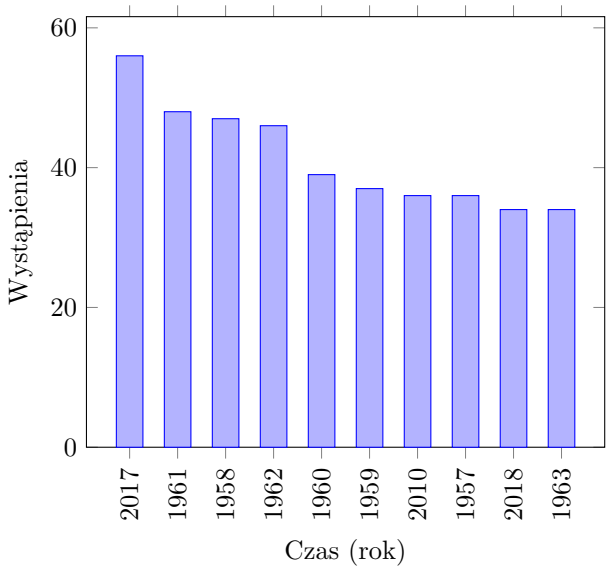
Rysunek 7: Najcześciej występujące lokalizacje dla grupy 'Doo-wop'

Dla grupy 'Doo-wop' obserwujemy głównie nawiązania do Nowego Jorku oraz Stanów Zjednoczonych. Ponadto pojawia się Bronx czyli być może jest to grupa z tego obszaru. Zaskakującymi są jednakże bardzo liczne wystąpienia Love, Love „ oraz Rock - które lokalizacjami nie są. Wykorzystany model do rozpoznawania lokalizacji (NLP) zwrócił takie wyniki, choć ich jakość wydaje się wątpliwa.



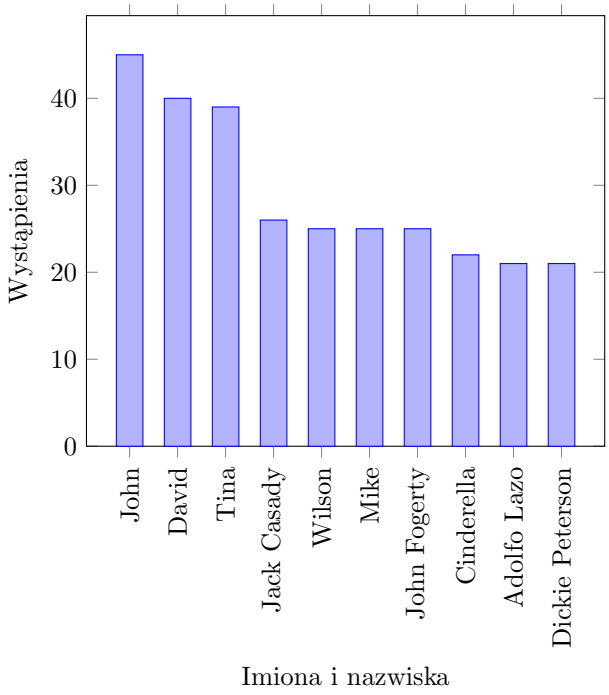
Rysunek 8: Najcześciej występujące lokalizacje dla grupy 'Electronic'

Dla 'Electronic' obserwujemy nawiązania do krajów, a w dwóch przypadkach do miast (Londyn oraz Los Angeles). Trend na tym wykresie jest najmniej zaskakujący ze wszystkich do tej pory scharakteryzowanych.

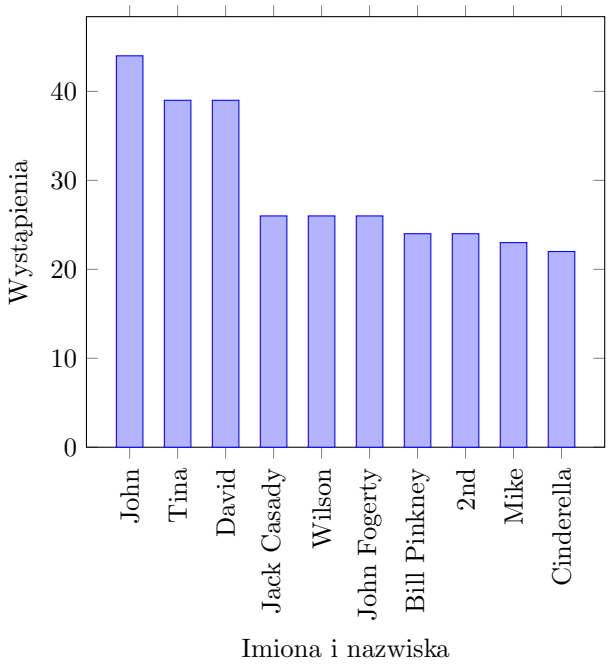


Rysunek 9: Najczęściej występujące lata dla grupy 'Doo-wop'

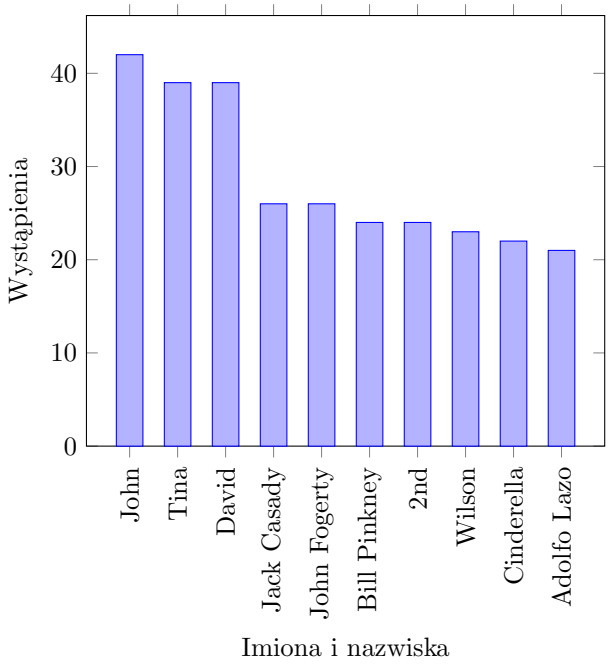
Analizując lata dla poszczególnych grup można zaobserwować bardzo zbliżone do siebie wystąpienia, ewentualnie minimalna przewaga któregoś z lat. W przypadku tej grupy mamy jednakże bardzo liczne wystąpienia lat .50 i .60 XX wieku - najprawdopodobniej szczyt popularności przypadał właśnie na ten okres.



Rysunek 10: Najczęściej występujące imiona i nazwiska dla grupy 'American blues rock'



Rysunek 11: Najcześciej występujące imiona i nazwiska dla grupy 'American blues'



Rysunek 12: Najcześciej występujące imiona i nazwiska dla grupy 'American instrumental'

Analizując jednocześnie odniesienia do tych trzech amerykańskich grup możemy zaobserwować ten sam wzorzec - jakie pierwsze trzy najliczniejsze imiona pojawiają się męskie 'John', 'David' oraz kobiece 'Tina'. Starając się zorientować nad poprawnością wyników, warto skupić się i zastanowić samemu - jakich artystów się kojarzy. Zdecydowanie te imiona będą najpopularniejsze (najprawdopodobniej).

5 Podsumowanie

Zgodnie z oczekiwaniami, w danych dotyczących analizowanych zespołów muzycznych występuje wiele ciekawych i zaskakujących trendów. Dzięki możliwości przeprowadzenia analizy z wykorzystaniem nowoczesnych narzędzi jakim jest *Apache Lucene*, *Apache Solr* czy *OpenNLP* można było zbadać te zależności. Był to również doskonały trening programistyczny w celu rozwiązania problemu analizy dużej grupy danych.