# Przeszukiwanie zespołów muzycznych w poszukiwaniu ciekawych zależności

Mateusz Kruszyna inf127252
Bartosz Górka inf127228
Jarosław Skrzypczak inf127265

## I. Wprowadzenie

Poniżej zostanie pokazane po krótce jak to zrobiliśmy i co nam wyszło ; )

## II. Pozyskanie danych

Do pozyskania danych posłużył nam Apache solr oraz Apache nutch. Z czternastu początkowych stron w domenie https://en.wikipedia.org/wiki/ oraz skonfigurowanie silnika szukającego tak, aby pomijał niepotrzebne linki (zewnętrzne strony, linkki do edycji, linki do mediawiki itp...) zostało pozykasne 1050 dokumnetów na temat zespołów muzycznych. Niestety 42 linki z powyższej kolekcji prowadzą do podkategorii, które to nie zostały uwzględnione w przeszukiwaniu.

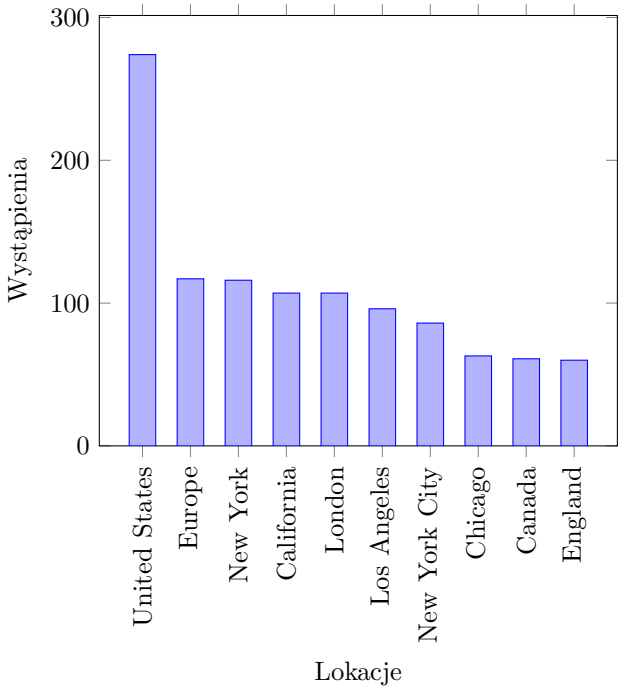
Drugim elementem było wgranie stworzonego indeksu do Apache Lucene, w którym dokonaliśmy analizy dokumentów. Z każdego dokumnetu pobraliśmy takie dane jak tytuł, listę najczęstszych lokalizacji, listę najczęściej występujących lat (tylko rok) oraz dopasowaliśmy najbardziej prawdobodobną kategorię do każdego dokumnetu z prędzej wygenerowanej listy możlwiych kategorii.

Krótko opisują cały mechanizm, to tytuł został wyciągnięty z linku do strony na wikipedii. Zawartość dokumnetu (plik html) został poddany tokenizacji. Z pomocą modułu do rozpoznawania lokalizacji zostały wybrane najczęściej występujące lokalizacje. Kategorie zostały rozbite na pojedyńcze słowa kluczowe i zostały zliczone sumaryczne wystąpienia każdego z tych słow w dokumencie. Pozyskiwanie dat (rok) z tekstu niestety nie okazało się dobre przy wykorzystaniu gotowego modelu z pakietu OpenNLP. W zamian tego zostały wykorzystane wyrażenia regularne do wykrycia prostych zapisów lat (format XXXX oraz 'XX dla lat 19XX).

Dla wszystkich powyższych danych zostały zastosowane proste metody eliminacji błędnych danych np. liczba 0001 jako rok, liczba 9988 jako rok, wyrażenie 'Random' jako lokalizacja, słowa 'by' lub 'genre' w nazwie kategorii.

Eksport danych do formatu csv został ograniczony do pierwszych 10 wystąpień danej zmiennej (czytelność wykresu), a także do danych, które występują więcej niż 1 raz w tekście.

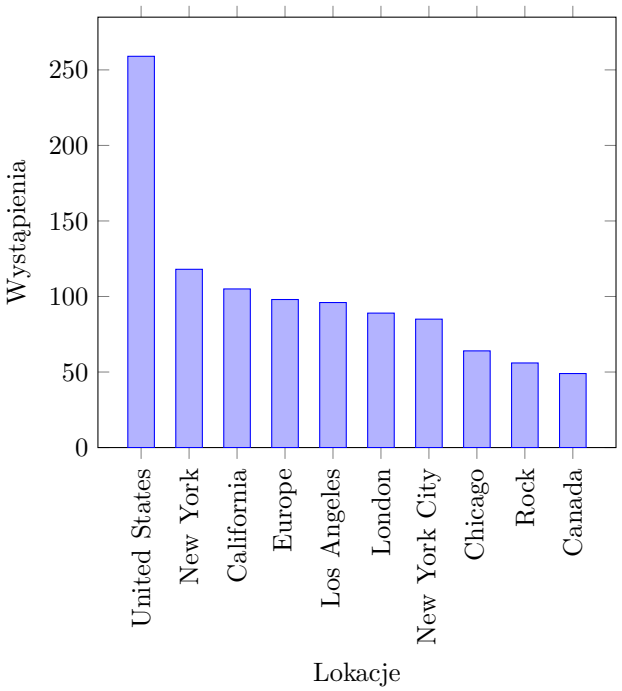
## III. Wykresy




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

## IV. Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.




Rysunek 2. Najcześciej występujące lokalizacje dla grupy 'American blues'