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 skonfigurowaniu 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 nie zostały uwzględnione w przeszukiwaniu, lecz strony znajdujące się w tych podkategoriach już tak.

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), listę najczęściej występujących imion i nazwisk oraz dopasowaliśmy najbardziej prawdobodobną kategorię do każdego dokumnetu z prędzej wygenerowanej listy możlwiych kategorii.

Cały mechanizm pozyskiwania danych do wykresów zostanie krótko opisany poniżej.

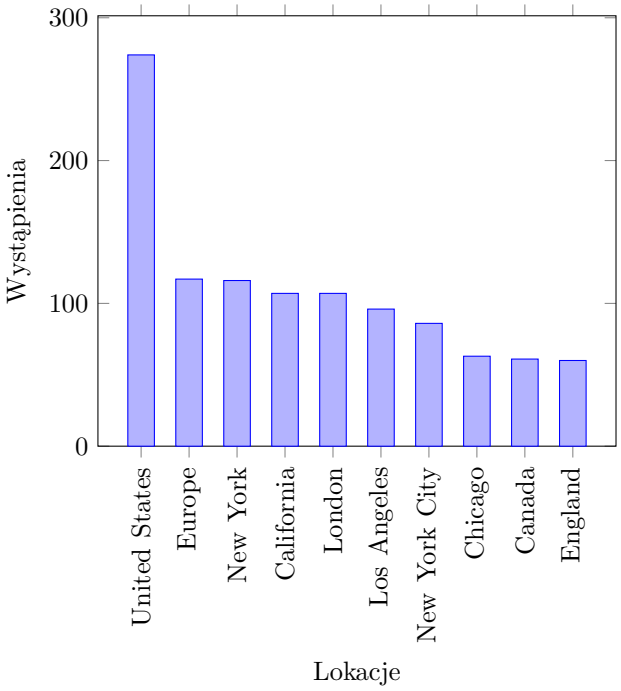
- Tytuł został wyciągnięty z linku do strony na wikipedii (końcówka linku).
- Zawartość dokumentu (plik html) został poddany tokenizacji, a następnie wykonane dalsze czynności
- Z pomocą modułu do rozpoznawania lokalizacji OpenNLP zostały wybrane najczęściej występujące lokalizacje. Pominięta została lokalizacja "Random", która naszym zdaniem była błednie rozpoznawana.
- Kategorie zostały wybrane z przeszukiwanej bazy, dzięki charakterystycznym schematom linków, które zawierały słowo kluczowe category:. Gdy w nazwie kategorii pojawiło się jedno ze słów zakazanych ("by","genre","navigational", "(genre)","musicians","nationality","body"), to taka nazwa nie została uznana za właściwą nazwę. Po takim zabiegu, aby wyszukać do jakiej kategorii należy dokument, dla każdego dokumentu kategorie zostały rozbite na pojedyńcze słowa kluczowe i zliczone sumaryczne wystąpienia każdego z tych słów w dokumencie.
- Pozyskiwanie dat (rok) z tekstu niestety nie okazało się dobre przy wykorzystaniu gotowego modelu z pakietu OpenNLP. W zamian tego zostały wykorzystane proste wyrażenia regularne do wykrycia dwóch zapisów zapisów lat (format XXXX oraz 'XX dla lat 19XX). Ograniczonych od dołu przez rok 1000 a od góry przez 2051.
- Wyszukiwanie imion i nazwisk zostało zaprogramowane dzięki przygotowanemu modułowi z OpenNLP. Tym razem także wyniki nie były jednoznaczne i wkradało się dużo różnych nie właściwych 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.

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.

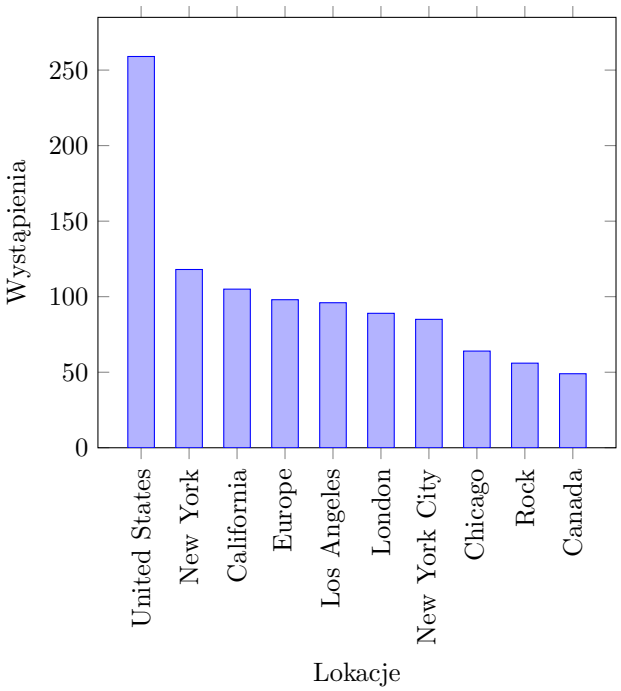
III. Wykresy

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



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