Jan Jędrzejewski
14 stycznia 2023

Projekt IUM

Pierwszy etap

Temat nr 3

“Może bylibyśmy w stanie wygenerować playlistę, która spodoba się kilku wybranym osobom jednocześnie? Coraz więcej osób używa Pozytywki podczas różnego rodzaju imprez i taka funkcjonalność byłaby hitem!”

Definicja problemu biznesowego

Celem zadania jest wygenerowanie playlisty, która będzie odpowiadała preferencjom kilku użytkowników jednocześnie. Dla podanych użytkowników należy wygenerować listę piosenek o zadanej długości, dzięki której użytkownicy będą częściej korzystać z naszej aplikacji i słuchać więcej muzyki.

Zdefiniowanie zadania

Stworzyć model rankingujący piosenki dla grupy użytkowników. Model ma wybierać utwory najbardziej zbliżone do zainteresowań i do utworów wcześniej przesłuchanych przez użytkowników.

Założenia i wymagania modelowania

Zbiory danych

Model ma dostęp do danych o użytkownikach, artystach, piosenkach i logach z sesji użytkowników. W sesjach są informacje o interakcjach użytkowników z utworami w systemie.

Wejście modelu

Podajemy listę użytkowników (id użytkowników) dla których chcemy stworzyć playlistę oraz długość wygenerowanej playlisty.

Wyjście

Lista piosenek (id piosenek) wybranych do playlisty.

Wymagania

Model ma wybierać piosenki pasujące do profilu użytkowników a nie w sposób losowy.

Analityczne kryterium sukcesu

Dla stworzonej przez model playlisty użytkownicy częściej polubią niż pominą piosenkę. Przykładowe kryterium sukcesu może być przedstawione wzorem:

$$\frac{(liczba\ polubień - liczba\ pominięć)}{liczbę\ rekomendacji} > \alpha - poziom\ skuteczności$$

Analiza danych

Ilość danych

W poniższej tabeli zestawiono porównanie liczby danych na początku i po otrzymaniu dodatkowych. Stworzenie użytecznego modelu byłoby niewykonalne przy tak niewielkim zbiorze użytkowników i sesji w porównaniu do bogatego zbioru utworów. Początkowo w informacjach sesji statystycznie była informacja tylko o 3 na 22 piosenki. Obecnie dla każdej piosence mamy średnio 10 zalogowanych informacji. 10 krotnie większa liczba użytkowników pozwoli na bardziej wszechstronne rozwiązanie pasujące do większej liczby odbiorców.

Nazwa zbioru danych	Początkowa liczba rekordów	Liczba rekordów
użytkownicy	50	500
artyści	1 667	1 667
utwory	22 412	22 412
sesje	3 198	285 114

Niepoprawne dane

W pierwotnym zbiorze danych, rekordy często zawierały wartości null. O ile przy nazwach lub id rekordu nie miałoby to większego wpływu dla zadania modelowania, to brak danych dla kluczy obcych np. id_artist, user_id, track_id jest bardzo problematyczny - tracimy informację o rekordzie bo nie wiemy dla jakiego użytkownika, utworu, artysty on się odnosi.

Wartości null były także obecne w kolumnach o informacji czy użytkownik wykupił “premium”, w popularności utworu oraz w wydarzeniach sesji.

W drugim zbiorze danych nie występują te anomalie. Poniżej tabela z porównaniem liczby niepoprawnych danych pomiędzy zbiorami danych:

Nazwa zbioru danych	Początkowa liczba błędnych rekordów/wszystkie	Obecnie
użytkownicy	4 / 50	0
artyści	83 / 1 667	0
utwory	4 146 / 22 412	0
sesje	481 / 3 198	0

Analiza rozkładów danych

W repozytorium zamieściliśmy pliki: `analyzeData.ipynb` z analizą pierwszego zbioru danych oraz `analyzeData2.ipynb` z drugim. Przyjrzeliśmy się uważnie liczbie konkretnych przykładów w zbiorach użytkowników, artystów i sesji, jak i rozkładom parametrów piosenek.

Rozkład użytkowników premium i nonpremium
Ulubione gatunki użytkowników
Gatunki artystów
Rodzaje akcji w sesjach
Atrybuty piosenek:

- popularity ma braki wartości dla około 63 i 75
- jeden nieproporcjonalnie długi utwór trwa ponad 68 minut (4120.258 sekund), dla porównania drugi najdłuższy trwa 23 minuty (1421.455 sekund)
- cykliczne braki utworów z lat
- instrumentalność dla większości utworów przyjmuje wartości bliskie zeru

Preprocessing

W ramach przetwarzania wstępnego przeprowadziliśmy następujące operacje:

- Odrzuciliśmy dane bardzo odstające, na przykład wyżej wspomniany utwór nieproporcjonalnie długi do pozostałych
- Dla daty wydania utworu odrzuciliśmy informacje o dniach i miesiącach wydania, zostawiliśmy sam rok wydania, ponieważ niektóre dane nie miały pozostałych informacji
- Znormalizowaliśmy dane parametryczne piosenek za pomocą biblioteki sklearn.preprocessing. Dzięki temu dane dla każdego z parametrów mają jednolite wartość i w modelu będą zawierały jednakową wagę.

W repozytorium zamieściliśmy plik `preprocessing.ipynb` w którym przedstawione zostały rozkłady parametrów utworów po normalizacji. Dane dla drugiego zbioru nie zawierają żadnych wartości null ani niejasnych id, dlatego nie ma potrzeby usuwania żadnych wierszy.

Realizacja zadania modelowania za pomocą dostarczonych danych

Dzięki danym o użytkownikach i o sesjach użytkowników będzie możliwe stworzenie modelu rankingującego piosenki na podstawie zainteresowań użytkownika. Szczegółowe dane o piosenkach pozwolą na odnalezienie wspólnych atrybutów ulubionych piosenek użytkowników. Dane o artystach pozwolą na dopasowanie gatunków muzycznych.