Temat

W ramach projektu wcielamy się w rolę analityka pracującego dla portalu „Pozytywka” - serwisu muzycznego, który swoim użytkownikom pozwala na odtwarzanie ulubionych utworów online. Praca na tym stanowisku nie jest łatwa - zadanie dostajemy w formie enigmatycznego opisu i to do nas należy doprecyzowanie szczegółów tak, aby dało się je zrealizować. To oczywiście wymaga zrozumienia problemu, przeanalizowania danych, czasami negocjacji z szefostwem. Same modele musimy skonstruować tak, aby gotowe były do wdrożenia produkcyjnego - pamiętając, że w przyszłości będą pojawiać się kolejne ich wersje, z którymi będziemy eksperymentować.

Jak każda szanująca się firma internetowa, Pozytywka zbiera dane dotyczące swojej działalności - są to:

- lista dostępnych artystów i utworów muzycznych,
- baza użytkowników,
- historia sesji użytkowników,
- techniczne informacje dot. poziomu cache dla poszczególnych utworów.

(analitycy mogą wnioskować o dostęp do tych informacji na potrzeby realizacji zadania)

Zadanie

“Fajnie byłoby rozszerzyć nasz serwis o generowanie popularnych playlist – zestawów pasujących do siebie utworów, których słuchaniem zainteresowane będzie wiele osób”.

Definicja Problemu Biznesowego

Kontekst:

Serwis muzyczny, który pozwala na odtwarzanie utworów online. Portal umożliwia tworzenie i edytowanie playlist.

Zadanie biznesowe:

Stworzenie playlist z popularnych utworów muzycznych, zbliżonych do siebie stylem, dostosowanych do jak największej grupy odbiorców.

Biznesowe kryterium sukcesu:

- Średni czas odsłuchu playlisty przez użytkowników jest dłuższy niż 70% sumarycznej długości wszystkich jej utworów – okres pomiaru wyników obejmuje miesiąc od momentu opublikowania playlisty.
- Utworzone playlisty osiągają 10% wskaźnik polubień utworów na niej.

Analityczne kryterium sukcesu:

$$y = c_1 \frac{\sum_{i=0}^{n} t_i}{n * T} + c_2 \frac{\sum_{i=0}^{n} \sum_{j=0}^{k} p_{ij}}{n * k}$$

y - zmienna celu

t_i - czas odsłuchu playlisty przez i-tego użytkownika

p_{ij} – ilość polubień utworów przez i-tego użytkownika na j-tej playliście

n – ilość użytkowników

k – liczba utworów na playliście

T – czas trwania wszystkich utworów w playliście

c_1, c_2 – stałe określające wagę poszczególnych składowych

Stopień korelacji analitycznego kryterium sukcesu z biznesowym kryterium sukcesu jest bardzo wysoki – dzięki temu spełnienie dowolnego kryterium powinno znacząco przybliżyć nas do spełnienia drugiego z nich.

Zadanie modelowania

Celem naszego projektu jest przyporządkowanie elementów z danej przestrzeni (na podstawie ich atrybutów) pod kątem ich relewantności do grupy użytkowników, dlatego zastosujemy w naszym projekcie zadanie modelowania jakim jest **rankingowanie**.

Ocena aktualnej sytuacji

W naszym posiadaniu znajdują się poniższe dane:

- Lista autorów razem z przypisanymi do nich gatunkami muzycznymi:
  - **Id: String**
  - Name: String
  - Genres: Array<String>
- Użytkownicy z podstawowymi danymi osobowymi i preferencjami muzycznymi
  - User_id: Int
  - Name: String
  - City: String
  - Street: String
  - Favourite_genres: Array<String>
  - Premium_user: Bool
- Tracki muzyczne razem z cechami dźwiękowymi oraz przypisanymi do nich artystami
  - Id: String
  - Name: String
  - **popularity: Int**
  - **duration_ms: Int**

    - **explicit: Bool**
    - **id_artist: String**
    - **release_date: Date,**
    - **danceability: Float**
    - **energy: Float,**
    - **key: Int**
    - **loudness: Float**
    - **speechiness: Float**
    - **acousticness: Float**
    - **instrumentalness: Float**
    - **liveness: Float**
    - **valence: Float**
    - **tempo: Float**
- Sesje użytkowników z timestampami wykonywanych przez nich akcji (play, skip, like)
    - **Session_id: Int**
    - **Timestamp: Date**
    - User_id: String
    - **Track_id: String**
    - **Event_type: enum{play, like, advertisment, skip}**

(pogrubione atrybuty zostaną użyte jako zmienne objaśniające)

Staramy się tworzyć listy dla jak największej grupy odbiorców, dlatego indywidualne preferencje użytkowników nie mają dla nas znaczenia. Głównym źródłem danych do wyznaczania popularności utworów będzie dla nas plik z sesjami użytkowników. Listy autorów analogicznie nie będą kluczowe.

W folderze plots przedstawiliśmy rozkłady kluczowych zmiennych dla naszego rozwiązania. Większość z nich jest na swój sposób przesuniętym, bądź zmodyfikowanym rozkładem normalnym. Niektóre natomiast bardzo znacząco od niego odbiegają – speechiness, instrumentalness, acousticness, key. Explicit jest zmienną rozkładu Bernoulliego.

Uważamy, że w danych występują błędy związane ze zmiennymi speechiness oraz instrumentalness – powinna występować silna ujemna korelacja, natomiast podczas analizy ustaliliśmy silnie dodatnią.

Naszym zdaniem dostarczone nam dane wystarczają do rozwiązania naszego problemu i zbudowania modelu rankingowania. Atrybuty silnie opisują charakter utworu, dzięki czemu możemy dosyć precyzyjnie ustalać ich ranking i tworzyć sensowne playlisty.

Lista wymagań, założeń i ograniczeń

Każda playlista musi znajdować się w przedziale czasowym <60, 180> minut – wiąże się to z tym, że przy długich sesjach użytkownik może odczuwać znudzenie i chęć zmiany gatunku muzycznego. Aktualnie bardzo mała grupa użytkowników ma więcej niż 3 godziny na słuchanie utworów.

Playlisty muszą posiadać unikalne utwory oraz powinny dynamicznie zmieniać autorów, aby użytkownik miał większą dywersyfikację słuchanych tracków.

Zakładamy, że dostarczone nam dane pochodzą z szerokiego przekroju wiekowego oraz preferencji gatunkowych użytkowników.

Gatunki muzyczne będziemy klasyfikować względem dostarczonych nam atrybutów utworów, wymienionych w opisie danych.

Polubienie piosenek jest jednorazowe i utrzymuje się przez cały okres użytkowania platformy.

Raz wygenerowana playlista nigdy nie ulegnie zmianie.

Zagrożenia dla projektu

Największą trudnością i jednocześnie zagrożeniem jakości rozwiązania będzie grupowanie utworów w gatunki i łączenie ich ze sobą w playlisty.

Istnieje też możliwość braku zainteresowania playlistami ze strony użytkowników, gdyż zbiory utworów generowane przez portal są mało zindywidualizowane – mają na celu trafienie do jak najszerszego grona odbiorców. Należy znaleźć złoty środek pomiędzy rozszerzaniem grona odbiorców, a indywidualizacją playlist.

Analiza kosztów i zysków z projektu

Kosztami projektu będzie przygotowanie odpowiedniego algorytmu przygotowującego playlisty oraz przygotowanie odpowiedniej dokumentacji. Nasza aplikacja będzie aktywowana na koniec każdego tygodnia – będzie to generowało dodatkowe koszty. Wdrożenie i utrzymanie aplikacji na serwerze będzie wymagało pracy małego zespołu. Zyskami będzie zwiększenie ilości przysłuchanych piosenek – im większą satysfakcję czerpie użytkownik z słuchania, tym więcej czasu na nie poświęca.

Określenie celów i etapów modelowania

Stworzenie playlisty o określonej długości i jak najlepszym dopasowaniu utworów przy użyciu narzędzi modelowania. Podobieństwo utworów określimy na podstawie cech utworów, zawartych w pliku tracks.jsonl. Etapy modelowania:

- Usuniemy duplikaty oraz błędne wartości
- Znormalizujemy dane
- Podzielimy dane na zbiór trenujący i testowy
- Wybieramy algorytm uczenia maszynowego: sztuczne sieci neuronowe
- Wygenerowanie potrzebnych danych do utworzenia zbioru treningowego
- Na podstawie otrzymanych danych trenujemy nasz model
- Testowanie modelu i określenie jego skuteczności
- Poprawianie modelu do uzyskania satysfakcjonujących wyników
- Wdrożenie modelu

Określenie podobieństwa gatunków

Dla każdej cechy utworu określimy jej ważność – przypiszemy stałą wartość opisująca, jak ważny jest dany parametr. Następnie na podstawie sum odległości euklidesowych porównamy do siebie utwory

i określimy ich podobieństwo. Do analizy i interpretacji otrzymanych danych użyjemy indeksu Silhouette.

Do grupowania utworów użyjemy sieci Kohonena – sztucznej sieci neuronowej realizującej uczenie nienadzorowane. Podany algorytm będzie trenowany do momentu otrzymania satysfakcjonujących wyników.