Inżynieria uczenia maszynowego - projekt

Tomasz Owienko Anna Schäfer

29.11.2023

1 Temat projektu

Temat projektu przekazany przez Klienta:

> *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!*

2 Problem biznesowy

TODO coś o samej Pozytywce w ramach kontekstu

Celem projektu jest realizacja funkcjonalności pozwalającej użytkownikom serwisu Pozytywka na generowanie playlist, których z których utwory podobać się będą wybranej grupie użytkowników. Taka funkcjonalność mogłaby być wykorzystywana do automatycznego układania playlist na imprezy w taki sposób, aby ich zawartość trafiała w gust jak największej części odbiorców. Implementacja takiej funkcjonalności ma zwiększyć zadowolenie użytkowników z jakości playlist odtwarzanych na imprezach, tym samym zwiększając ich zadowolenie z użytkowania portalu.

Biznesowe kryterium sukcesu

TODO możliwości - wybrać jakiś konkret:

- Dla $k\%$ użytkowników, playlista odtwarzana przez co najmniej $X\%$ jej czasu trwania w ciągu jednej z najbliższych N sesji użytkownika
- Dla $k\%$ użytkowników, podczas odtwarzania playlisty przez co najmniej $X\%$ jej czasu trwania pomijane jest co najwyżej $Y\%$ utworów
- Coś jeszcze?

2.1 Założenia

- Playlisty generowane będą na podstawie profili oraz historii sesji nie więcej niż N użytkowników
  TODO ile to N
- Playlisty w większości przypadków użycia nie będą wykorzystywane wielokrotnie
- Dobór kolejności utworów na playliście nie jest przedmiotem zadania
- Dostęp do playlisty mają wszyscy użytkownicy, których profile i historia sesji były uwzględnione przy jej generowaniu / tylko użytkownik, który utworzył playlistę

2.2 Pożądane cechy rozwiązania

- Playlista może być wygenerowana w bardzo krótkim czasie
- Funkcjonalność zachowuje się poprawnie dla nowo dodanych użytkowników oraz utworów
  TODO cold start - czy istotny?
- W ocenianiu gustu muzycznego poszczególnych użytkowników większe znaczenie powinny mieć niedawno odtwarzane utwory

3 Zadanie modelowania

Projekt zakłada zamodelowanie problemu jako zadanie generowania rekomendacji. Planowane jest zastosowanie podejścia *collaborative filtering*, które (w kontekście zadania) opiera się na wyszukiwaniu użytkowników podobnych do rozpatrywanych i generowania rekomendacji w oparciu o ich historie sesji. Do realizacji podejścia *collaborative filtering* zastosowana zostanie technika rozkładu macierzy interakcji między użytkownikami, a utworami. Przewidziane jest porównanie jakości modelu korzystającego z macierzy *feedbacku niejawnego* (użytkownik X odtworzył utwór Y), oraz *feedbacku jawnego* (użytkownik X wystawił utworowi Y ocenę Z).

Podejście *collaborative filtering* pozwala na generowanie rekomendacji dla pojedynczego użytkownika. Aby dostosować je do problemu, generowanie rekomendacji dla wielu użytkowników jednocześnie zamodelowane będzie jako:

1. Wygenerowanie bardzo dużej liczby rekomendacji dla każdego z użytkowników wraz z oceną podobieństwa poszczególnych utworów do gustu muzycznego użytkownika (liczba rekomendacji proporcjonalna do liczby użytkowników)
2. Znormalizowanie ocen podobieństwa do zakresu $< 0, 1 >$
3. Wyznacznie zbioru utworów, które pojawiły się w rekomendacjach wszystkich użytkowników
4. Wybranie tych utworów, dla których iloczyn ich ocen dla wszystkich użytkowników jest największy

Istotnym problemem w rozważanym zadaniu jest tzw. *cold-start*, czyli zachowanie modelu dla użytkowników bądź utworów, na których model nie był trenowany. Tradycyjne podejścia rozkładu macierzy interakcji takie jak *Funk MF* czy *SVD++* nie przewidują występowania takich sytuacji i spisują się słabo w scenariuszach *cold-start*. W rozwiązaniu zostanie wykorzystany model LightFM, który rozwiązuje problem *cold-start* przez zastosowanie metadanych (atrybutów) do opisywania zarówno użytkowników, jak i utworów. Przykładowo, nowy użytkownik dodany do systemu nie posiada historii sesji, ale jest znany jego wiek, płeć, czy chociażby zadeklarowane preferencje, co do gatunków muzycznych - dane te można wykorzystać do wyszukiwania podobieństw względem innych użytkowników.

4 Analiza dostarczonych danych

TODO kolejne iteracja zbierania danych od Klienta, dodatkowe wymagania od zadania (liczba użytkowników, więcej historii sesji)

TODO - czy mamy opisywać cały feature engineering?

Z perspektywy wytestowania podejścia opartego na *jawnym feedbacku* należy rozważyć sposób zamodelowania takiej informacji - nie jest ona dana bezpośrednio w zbiorze danych. Jednym z możliwych podejść może być wyznaczenie domniemanej oceny wystawionej przez użytkownika danemu utworowi na podstawie:

- Częstotliwości odtwarzania danego utworu
- Częstotliwości występowania zdarzenia `like`
- Częstotliwości występowania zdarzenia `skip`