10.8. Tworzenie profilu użytkownika

Jeśli lubisz Jamesa Bonda, to może również spodoba ci się... To jest prosty przykład, który nie wymaga żadnego profilu użytkownika. Z modelem LDA lub po prostu wektorami cech, o których była mowa wcześniej, znajdziesz wektor filmu o Jamesie Bondzie, a następnie zlokalizujesz filmy z podobnymi wektorami. Jeśli jednak oferujesz spersonalizowane rekomendacje, musisz utworzyć profil użytkownika obejmujący wszystkie filmy, które podobają się użytkownikowi. Spójrzmy, jak to zrobić za pomocą LDA, a potem TF-IDF.

10.8.1. Tworzenie profilu użytkownika za pomocą LDA

Podczas tworzenia funkcji spersonalizowanych rekomendacji należy przyjrzeć się pełnej liście elementów, które podobają się użytkownikowi, i zwrócić inne elementy, które użytkownik może również polubić. W prawdziwym życiu funkcja ta nie jest indywidualna – użytkownik może nie być jednoznacznie zdefiniowany przez skonsumowane elementy (nawet jeśli większość systemów rekomendacji prawdopodobnie odpowie w ten sam sposób). Ale nadal jest to o wiele bardziej spersonalizowane niż lista najlepszych elementów.

To, co można zrobić, to dokonać iteracji po każdym z elementów, które użytkownik lubi, i dla każdego znaleźć podobne elementy. Gdy pojawi się lista, uporządkuj ją według iloczynu podobieństwa i ocen użytkowników z oryginalnej listy skonsumowanych elementów. Bardziej formalnie, dla aktywnego użytkownika możesz wykonać następujące czynności:

- Pobierz wszystkie elementy (CI) skonsumowane przez aktywnego użytkownika.
- Dla każdego elementu I w CI:
  - Znajdź podobne elementy za pomocą modelu LDA.
  - Oblicz ocenę na podstawie podobieństwa i ocen aktywnych użytkowników.
- Poukładaj elementy według oceny.
- Poukładaj według relewancji (jeśli masz jakieś dane na ten temat).

Alternatywnym sposobem, opisanym przez Jobina Wilsona i in., jest utworzenie macierzy LDA użytkownika, a następnie znalezienie podobnych elementów[55]. Interesujące również może być zobaczenie, jakby to działało. Ocena wygląda bardzo dobrze.

10.8.2. Tworzenie profilu użytkownika za pomocą TF-IDF

Przy użyciu opisanych wcześniej wektorów dotyczących tagów i faktów, można w inny sposób utworzyć profil użytkownika. Możesz agregować macierze elementów, które użytkownik lubi i odejmować elementy, których użytkownik nie lubi. Spójrz na wnikliwe informacje w tabeli 10.3.

---

[55] Jobin Wilson i in., Improving Collaborative Filtering Based Recommenders Using Topic Modelling, Luty 2014. Streszczenie pod adresem https://arxiv.org/abs/1402.6238.

Tabela 10.3 Wektor reprezentujący kilka filmów

	Występuje Ben Affleck	Akcja	Film przygodowy	Komedia	Wybuchy[a]
BvS	1	1	1	0	5
Valentine's Day	1	0	0	1	0
Poszukiwacze zaginionej Arki	0	1	1	1	2
La La Land	0	0	0	1	0

[a] Te liczby są wymyślone.

Jeśli masz użytkownika (nie robisz tutaj wspólnej filtracji, więc wystarczy spojrzeć na jednego), który dał 5 gwiazdek *Poszukiwaczom zaginionej arki* (kto nie kocha tego filmu?) i 3 gwiazdki *La La Land*, możesz teraz utworzyć profil użytkownika, mnożąc oceny w wektorze filmu i dodając je, jak pokazano w tabeli 10.4.

Tabela 10.4. Mnożenie ocen użytkowników w wektorach filmów i dodawanie każdego elementu w celu utworzenia profilu użytkownika

	Występuje Ben Affleck	Akcja	Film przygodowy	Komedia	Wybuchy[a]
Poszukiwacze zaginionej Arki	0	5	5	5	10
La La Land	0	0	0	3	0
Profil użytkownika	0	5	5	8	10

[a] Te liczby są wymyślone.

Teraz można użyć tej macierzy, aby znaleźć podobną zawartość dla użytkownika. Prawdopodobnie należy znormalizować wartości tak, aby były w tej samej skali co filmy, a liczba wybuchów powinna być prawdopodobnie pomniejszona w filmie, który zawiera albo nie zawiera wybuchów. W takim przypadku profil użytkownika wygląda jak w tabeli 10.5.

Tabela 10.5. Mnożenie ocen użytkownika w wektorach filmów i dodawanie każdego elementu, aby utworzyć profil użytkownika

	Występuje Ben Affleck	Akcja	Film przygodowy	Komedia	Wybuchy
Profil użytkownika	0	5	5	8	5

Uważam, że ta procedura jest bardziej wiarygodna dla filmów. Dobrą rzeczą jest to, że można szukać filmów, które mają wszystkie aspekty tego, co podoba się użytkownikowi (jeśli oczywiście uchwycisz odpowiednie tagi i fakty), ale chociaż lubię czekoladę

i lasagne, to nie oznacza, że lubię je razem. To samo można powiedzieć o wielu innych atrybutach, jakie można wymyślić zarówno dla żywności, jak i dla filmów. Można tego użyć, aby zobaczyć, czy użytkownik lubi komedie bardziej niż filmy akcji i przygodowe.

W części analitycznej witryny MovieGEEKs przedstawiłem gust użytkownika, przeglądając filmy ocenione przez użytkowników i zsumowanie ocen dla każdego gatunku, jak pokazano na poniższym listingu. Kod można wyświetlić w pliku /analytics/views.py.

Listing 10.7. Wyodrębnianie gustu z ocen

```
for movie in movies:
    id = movie.movie_id

    rating = ratings[id]

    r = rating.rating
    sum_rating += r
    movie_dtos.append(MovieDto(id, movie.title, r))
    for genre in movie.genres.all():

        if genre.name in genres_ratings.keys():
            genres_ratings[genre.name] += r - user_avg
            genres_count[genre.name] += 1

max_value = max(genres_ratings.values())
max_value = max(max_value, 1)
max_count = max(genres_count.values())
max_count = max(max_count, 1)

genres = []
for key, value in genres_ratings.items():
    genres.append((key, 'rating', value/max_value))
    genres.append((key, 'count', genres_count[key]/max_count))
```

Obliczone dla każdego filmu, który użytkownik ocenia

Pobiera ocenę

Iteruje po każdym gatunku filmu i buduje słownik z nazwami gatunków jako kluczami i sumą ocen jako wartościami

Upewnia się, że nie jest mniejsze niż jeden

Wyszukuje wartości maksymalne

Upewnia się, że nie jest poniżej jednego

Normalizuje wartości

Wyszukuje wartości maksymalne

Ten kod jest używany do tworzenia wykresów, takich jak ten na rysunku 10.18 dla użytkownika 100 (http://localhost:8001/analytics/user/100/).

10.9. Rekomendacje oparte na treści na witrynie MovieGEEKs

Jak wspomniano kilka razy, będziemy przechodzić przez implementację i wykorzystamy budowanie modelu LDA opartego na treści. Najpierw musimy powiedzieć kilka słów na temat pobierania danych.

10.9.1. Pobieranie danych

Zbiór danych, z którego korzystasz, nie zawiera opisów filmów, więc znów jesteś na łasce www.themoviedb.org, aby pobrać dane. W katalogu głównym kodu towarzyszącego książce znajduje się skrypt o nazwie populate_sample_of_descriptions.py, który pobiera opis najnowszych filmów. Przykład tego, co można pobierać, pokazano na rysunku 10.19[56].

[56] Zobacz http://MNG.bz/1UXq.