Systemy rekomendacyjne ogólna idea

Mamy info o użytkownikach i produktach mamy tablice dla ułatwienia oceny 1-5, generalnie w realnych warunkach jest tego dużo więcej



CF User-based

Żeby było prościej zrozumieć dodajmy trochę określeń na realnych danych

Musimy teraz przeanalizować podobieństwa ocen

Po pierwsze nie wszyscy ocenili wszystko naszym celem będzie propozycja uzupełnienia pustych ocen

Po drugie ile razy poszczególni użytkownicy ocenili dany produkt i jest to min 2

Skupmy się na początke tylko na jednym użytkowniku



Chcemy wskazać jakie oceny Ann dałaby filmom Pul Ficiotn i Ant Man. Przykładowo będziemy mogli skategoryzować filmy i zakładamy że grupowanie odbyłoby się bez dostępu do danych i jedne fil będzie mógł wpadać do kilku grup.

Ania zgodziła się z Zosią co do oceny Zielonej mili i z Hanią co do Iron man po za tymi ocenami różniły się nieznacznie od pozostałych. Największa różnica jest z Kasią pod względem Up. Naszym celem będzie określenie prawdopodobieństwa użytkowników pod względem ich ocen.

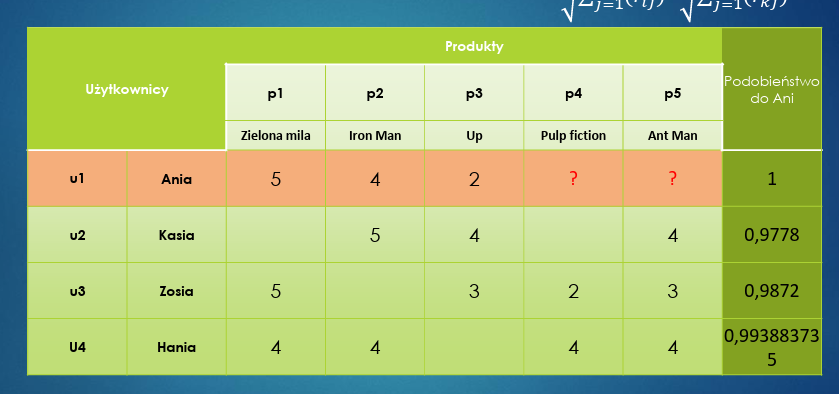
Procedura:

Wybieramy dwóch użytkowników

Niech rij będzie oceną wystawioną przez i-tego usera j-temu produktowi gdzie j =1,..,w to indeksy kolejnych produktów ocenionych przez dwóch użytkowników

Analogicznie rkj to ocena k-tego usera dla j-tego itemu

CF- miary prawdopodobieństwa (kartka z notatką) dobrze jakby wartości byłyby ograniczone, adaptacja przebywa 1/odległość w innych więc w sumie nie jest to tak dobre jak cosinusowa generalnie chcemy uzyskać wyniki im bliżej 1 lub 0



Wracając do naszego usera użyjemy miary cosinusowej

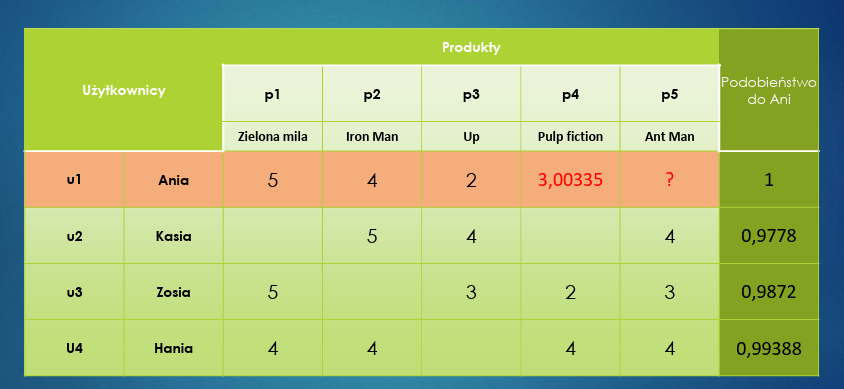
Liczymy podobieństwo Ani do Kasi:

Wybieramy filmy wspólnie ocenione: Iron Man i Up

Liczymy odległość cosinusową: (4\*5+2\*4) /(pierwiastek z 4\*4+2\*2) (pierwiastek z 5\*5+4\*4 ) = k

Najbardziej podobna jest Hania jest Winnerem

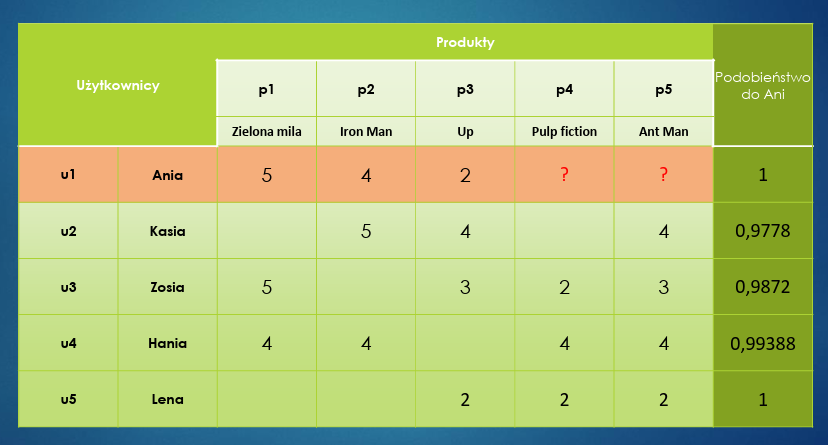
Kolejnym krokiem jest uzupełnienie pustych ocen : suma ważona ocen pozostałych użytkowników (wzór w notkach) sumowanie osbywa się po wszystkich użytkownikach którzy ocenili j-ty przedmiot wagami są współczynniki prawdopodobieństwa



Przewidzenie oceny Ani dla Pulp fiction to bierzemy odpowiedzi od Zosi i Hani. Zosia wystawiła 2 i bierzemy jako wagę jej stopień przynależności do Ani dodajemy Hanię i wychodzi : (2\*0,9872 + 4\*0,99388)/0,9872+0,99388. Dla Ant Man wyszło 3,666 tutaj bierzemy trzech userów do obliczeń

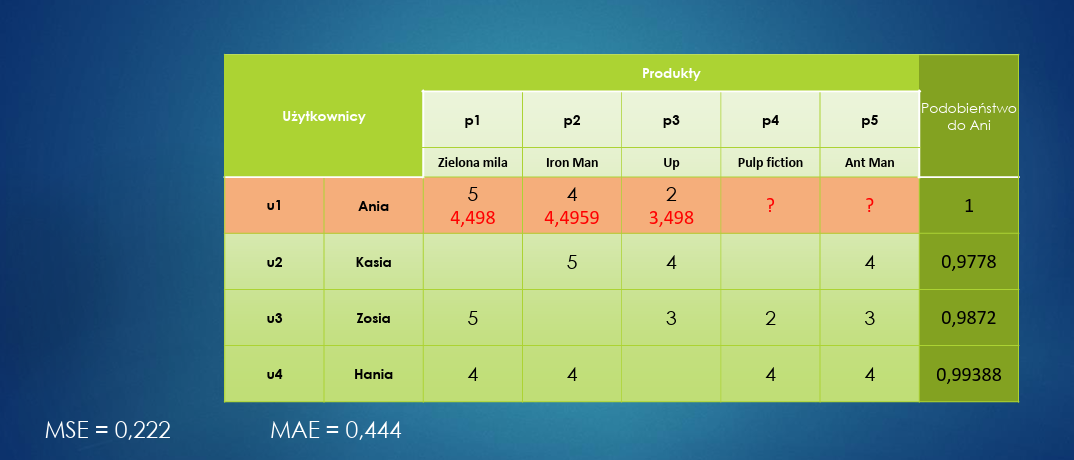
Teraz możemy przejść do rekomendacji i zaproponowalibyśmy Ant Mana ale najczęściej proponujemy n propozycji z największym wynikiem

Problem pojawia się w tym że bazy są ogromne i aby to zminimalizować to obliczamy prawdopodobieństwo bazując na k userach a nie na wszystkich czyli w następnych krokach nie będziemy używać wszystkich ale na podstawie podobieństwa wybierzemy k podobnych i na nich będziemy liczyć średnie ważone. Ograniczając się do k musimy wziąć pod uwagę to że można ograniczyć ilość produktów które możemy zaproponować.



Kolejnym problemem jest wiarygodność użytkowników Hania oceniła wszystkie filmy tak samo jest to typowy problem że na przykład oceniamy 1 albo 5 więc warto wprowadzić więcej czynników niż tylko rangi.

Innym problemem jest brak wspólnych ocen jest nowy użytkownik Lena i oceniła tylko Up wraz z Anią i prawdopodobieństwo wyszło 1 często ustala się próg ile wspólnie muszą ocenić filmów.



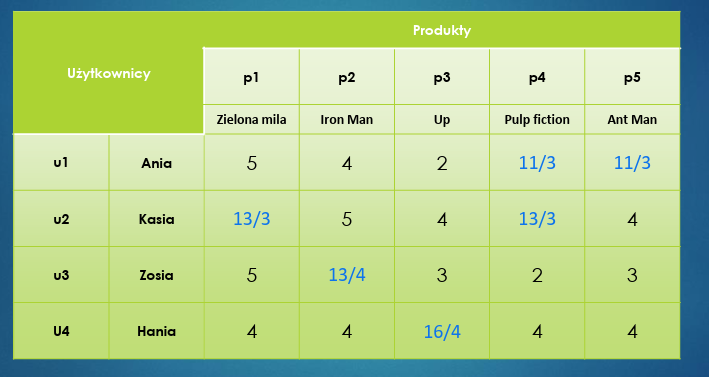
Stosując to podejście możemy sprawdzić jak system przewiduje już ocenione produkty i będzie to jakaś miara naszego sytemu

Miary MSE i MAE to dokładność estymacji ocen nie całego systemu biorąc pod uwagę że błąd mógłby mieć max 4 więc MSE wskazuje na słabość systemu

Czasem żeby uwzględnić to zróżnicowanie ocen u userów to uwzględnia się średnią ze wszystkich ocen użytkownika (wzór w pdfie oznaczony \*)

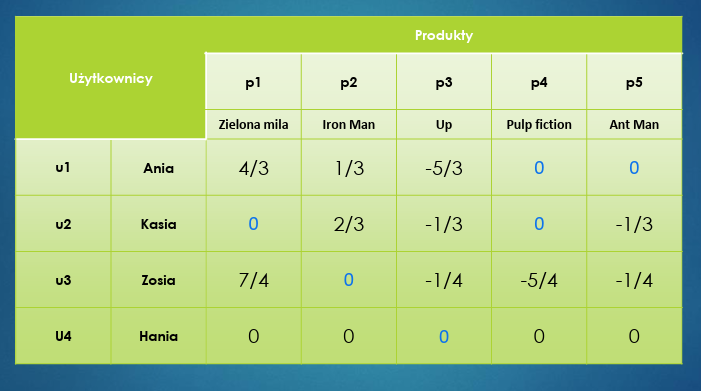
Podsumowując procedura CF odbywa się w następujący sposób:

1. Tworzona jest macierz ocen (wiersze to użytkownicy a kolumny produkty)
2. Obliczane są prawdopodobieństwa między userami
3. Wyszukujemy K użytkowników o największym współczynniku prawdopodobieństwa dla naszego i-tego usera. Uwzględniamy czy zostało spełnione kryterium o minimalnej liczbie wspólnych ocen
4. Generujemy prognozy produktów
5. Wybieramy produkty nie ocenione wcześniej przez i0tego usera
6. Wskazujemy N produktów o najwyższych szacowanych ocenach



Można spotkać się z podejściem gdzie pierwsze tabele są porównywane w całości czyli wymagane jest uzupełnienie brakujących ocen na przykład średnia z wiersza nie zaleca się dawanie liczby spoza zakresu

W przypadku użycia miary prawdopodobieństwa cosinusowego często ta wartość średnia jest odejmowana aby uzyskać miary zcentralizowane wokół 0 nazywa się to zcentralizowana miara cosinusowa i jest równoważna współczynnikowi korelacji Pearsona

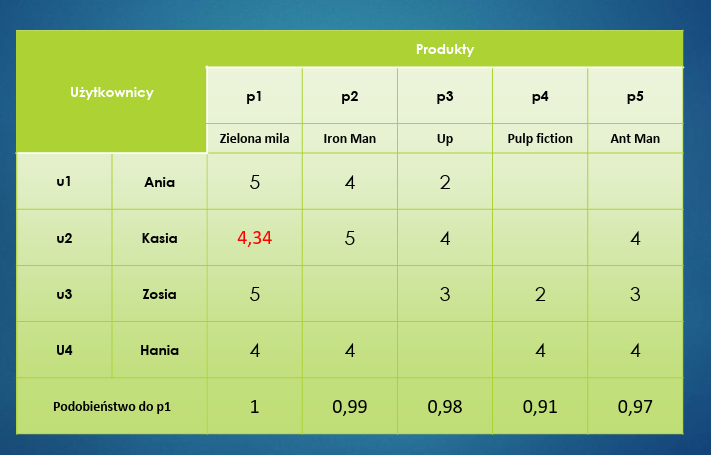


Collaborative filtering- Item based

To podejście jest analogicznie tylko teraz szukamy rzeczy ocenionych w ten sam sposób wcześniej szukaliśmy podobnych userów



Wykorzystując prawdopodobieństwo cosinusowe chcemy ocenić jaką ocenę wystawiłaby Kasia filmowi Zielona mila. Musimy policzyć prawdopodobieństwo każdej kolumny. Prawdopodobieństwo między Zielona mila a Iron Man będzie wynosiło: (5\*4+4\*4)/(pierwiastek5\*5+4\*4) pierwiastek (4\*4+4\*4) = 0,99



Samą ocenę Zielonej mili wystawimy na podstawie Iron Man, Up oraz Ant Man i mamy: 5\*0,99+4\*0,98+4\*0,97/0,99+0,98+0,97=4,34

Oblicz rangi dla Pul fiction i Ant Man i który zostanie zaproponowany Ani. To odbywa się analogicznie- obliczamy dla wszystkich wartości, wybieramy n najwyższych rang pośród tych nieocenionych przez Anię. Często to podejście nazywane jest k-nn dla rekomendacji.