

Wprowadzenie. Content-based recommenders

Systemy Rekomendacyjne 2021/2022

Plan wykładów

1. Wprowadzenie. Content-based recommenders
2. Collaborative filtering
3. Multi-armed bandits
4. Podział użytkowników na grupy zainteresowań (segmentacja)
5. Rekomendacje grafowe [TBD]
6. Algorytmy konwersacyjne [TBD]
7. Zagadnienia praktyczne – przechowywanie i przetwarzanie danych, systemy rekomendacyjne na produkcji

Plan laboratoriów

1. Content-based recommenders
2. Collaborative filtering
3. Multi-armed bandits
4. Podział użytkowników na grupy zainteresowań (segmentacja)
5. Rekomendacje grafowe [TBD]
6. Algorytmy konwersacyjne [TBD]
7. Wystawienie ocen

Oceny

- Za każde laboratorium można otrzymać 1 punkt
 - Laboratorium oceniane jest na tych samych lub następnych zajęciach
- Ocena z laboratoriów (i równocześnie ocena końcowa) wystawiana jest na podstawie sumy punktów, zgodnie ze skalą AGH

Czym są
systemy rekomendacyjne?

Nagrodzone Emmy wciągające seriale i programy

Wybrane dla: Piotr

Moja lista

Nowości

Filmy, seriale i programy

- Polecane
- Jak zostać gwiazdą
 - Jak zostać królem
 - Dragons' Den - jak zostać milion...

- Strona główna
- Anime
 - Dramaty
 - Kino niezależne
 - Thrillery
 - Filmy i seriale historyczne
 - Horror
 - Filmy rodzinne i dla dzieci
 - Seriale

Not Sure What to Watch?

Choose Play Something and we'll pick things for you to watch based on your tastes.

TV Dramas




PLAYLISTA

Odkryj w tym tygodniu

Tygodniowa dawka świeżych brzmień. Nowości wybrane specjalnie dla Ciebie. Aktualizacja co poniedziałek.

Spotify • 1 polubienie • 30 utworów, 2 godz. 25 min











PLAYLISTA

Daydreaming Radio

Wraz z Radiohead, UNKLE, Damon Albarn, Jeff Buckley i innymi

Spotify • 264 polubienia • 50 utworów, 3 godz. 47 min



#	TYTUŁ	ALBUM
1	 Daydreaming Radiohead	A Moon Shaped Pool
2	 Ingenue Atoms For Peace	AMOK
3	 Suspirium Thom Yorke	Suspiria (Music for the Luca Guadagnino Film)
4	 By This River - 2004 Digital Remaster Brian Eno	Before And After Science
5	 I'm So Tired Fugazi	Instrument Soundtrack

Przygotowano dla Piotr

Im więcej słuchasz, tym lepsze kawałki Ci polecamy



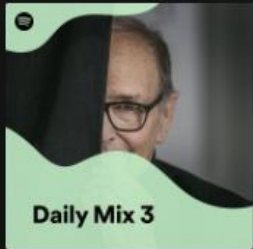
Daily Mix 1

Herbert von Karajan,
Hallé, Jacqueline du Pré ...



Daily Mix 2

Dinu Lipatti, Hilary Hahn,
Radu Lupu i więcej



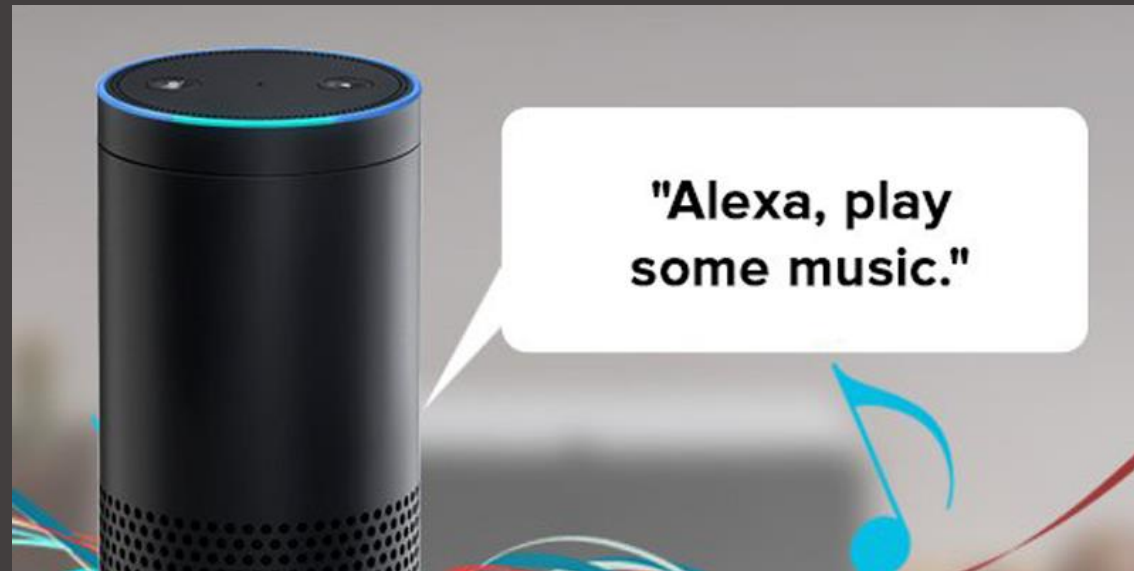
Daily Mix 3

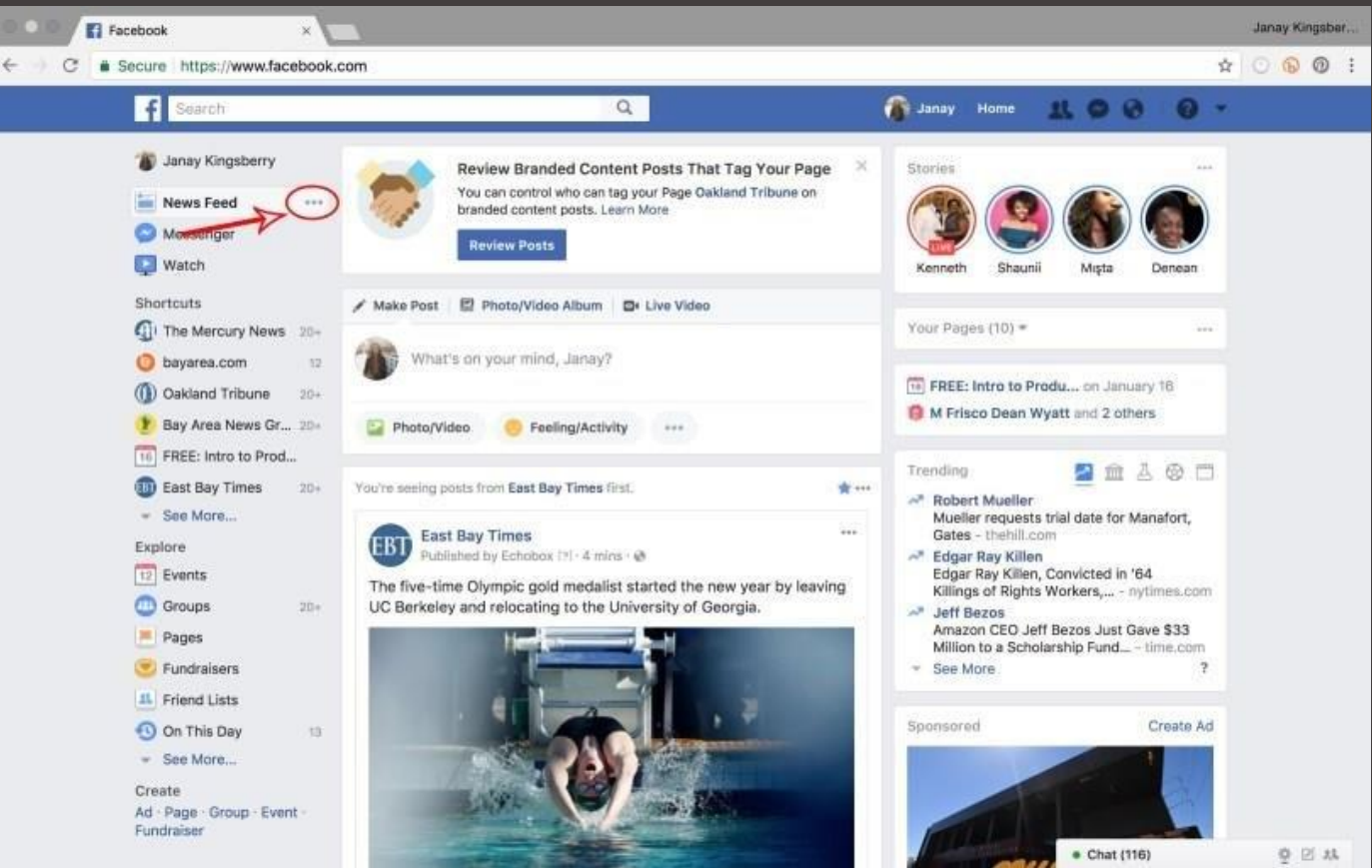
Ennio Morricone, Bear
McCreary, Kevin MacLe...



Daily Mix 4

Víctor Pellegrini, John
Williams, Orchestra Del...





Osoby, które możesz znać

[Pokaż wszystkich](#)

Alice 4 wspólnych znajomych Dodaj do znajomych Usuń	Bob 2 wspólnych znajomych Dodaj do znajomych Usuń	Carol 17 wspólnych znajomych Dodaj do znajomych Usuń	Dave 8 wspólnych znajomych Dodaj do znajomych Usuń

Proponowane dla Ciebie

[Zobacz więcej](#)

Grupy, które mogą Cię zainteresować

#błm

[←](#)[→](#)

Praca dla programistów i programistek powyżej 20000zł, 30000zł i 40000zł
24 tys. członków grupy
30 znajomych jest członkami grupy

[Dołącz do grupy](#)



Naukowcy prognozują zmiany klimatu w Polsce. Czeka nas galopujący wzrost temperatury



Kierowca zatrzymany po obławie. Śmiertelnie potrącił 4-latkę



O jej przemówieniu mówi cała Polska. Teraz zabrała głos



Medycy ze Szpitala Narodowego złożyli zawiadomienie do prokuratury



Bąkiewicz dostał trzy mln zł od państwa, m.in. na sprzęt



Wymiana zdań między dziennikarzem TVN a postem PiS w sprawie TK



Nieoficjalnie: w przyszłym tygodniu rząd przedstawi nowych ministrów

WYBRANE DLA CIEBIE



Straż Graniczna zamieściła nagranie z cudzoziemcami na granicy



Historyczna klęska Trójki. Zaskakujący komentarz Marka Suskiego

WIADOMOŚCI



Rozpędzone auto wjechało w 4-latkę. Złapano kierowcę

Były agent z Korei Północnej przerywa milczenie "To kwestia dni. Tusk złoży propozycję" "Internetowa apokalipsa". Burza uderzy w Ziemię Jutro spadnie krupa śnieżna i śnieg w Polsce 7-latek terroryzuje rówieśników w szkole Gowin pisze list do wyborców Spicie dziennikarza i posła PiS ws. wyroku TK Wir polarny jest zakłócany, co zwiastuje zimą Za kilka dni możemy poznać nowych ministrów

- Miał władzę i był lojalnym, komunistycznym sługą, ale i tak musiał uciec z Korei Północnej w 2014 r.
- Od tego czasu mieszka w Seulu i pracuje dla wywiadu Korei Południowej
- Teraz opowiada BBC o tym, co musiał zrobić, by spełnić oczekiwania północnokoreańskich przywódców
- Dla nich zlecał zabójstwa, a także stworzył profesjonalne laboratorium metaamfetaminy
- Więcej takich tematów znajdziesz na stronie głównej Onetu



Więzenia w Korei Południowej były kiedyś wypełnione szpiegami aresztowanymi przez dziesięciolecia za różnego rodzaju prace szpiegowskie na rzecz Korei Północnej. Jednak od 2017 r. jest ich mniej, bo kraj Kim Dzong Una stawia na nowe technologie. Uciekinierzy z Pjongjangu ostrzegają, że Korea Północna ma armię sześciu tys. wykwalifikowanych hakerów. I jest to prawdopodobne, bo to północnokoreańska grupa Lazarus miała stać za cyberatakami, który sparaliżował część NHS (National Health Service - red.) i innych organizacji na całym świecie w 2017 r.

- Polecamy: Koszmar złapanych uciekinierów. Kim Dzong Un przykręcił śrubę



Wypadek miał miejsce w niedzielę na skrzyżowaniu ulic Armii Polskiej i 30 stycznia. Najpierw doszło tam do zderzenia dwóch samochodów osobowych – chevroleta i BMW. Pierwszy z nich odbił się i potrącił 4-letniego chłopca. Na skutek odniesionych obrażeń dziecko zmarło.

ZOBACZ RÓWNIEŻ



Kierowca śmiertelnie potrącił 4-latkę i uciekł. Policja ma "swoje typy"



Tragedia w Gorzowie Wielkopolskim. Rozpędzone auto wjechało w 4-latkę

– Bezpośrednio po wypadku kierowca chevroleta pieszo oddalił się z miejsca zdarzenia. Obecnie poszukują go policjanci ze wszystkich pionów gorzowskiej komendy oraz z oddziałów prewencji. Nie wykluczamy, że zbiegł, gdyż mógł prowadzić auto pod wpływem alkoholu, bądź narkotyków – tłumaczył mediom w poniedziałek rano Marcin Maludy, rzecznik lubuskiej policji.

Google

recommender |

- recommender **systems**
- recommender **systems machine learning**
- recommender **meaning**
- recommender **synonym**
- recommender **system is an example of**
- recommender **system python**
- recommender **system algorithms**
- Recommender Systems Handbook
- recommender **common app**
- recommender **engine**

Google Search

I'm Feeling Lucky

Report inappropriate predictions

Google

spotify

All Images Videos Maps News More

Tools

About 350,000,000 results (0.57 seconds)

Ad · <https://www.spotify.com/>

Spotify® Premium - Wypróbuj miesiąc za darmo

Twoi ulubieni wykonawcy i muzyka bez przerw. Potem 19,99 zł / mies. Obowiązują Warunki. Odkryj świetnych muzyków. Playlisty na każdą chwilę.

Śłuchaj gratis w Spotify®

Odtwarzaj utwory, które kochasz za darmo. Zarejestruj się już dziś

Spotify® dla studentów

Spotify® Premium za 50% ceny
1 miesiąc za darmo. 9,99 zł/mies.

Spotify® dla rodziny

Podaruj bliskim muzykę bez końca.
1 miesiąc za darmo. 29,99 zł/mies.

Spotify® Premium Duo

2 ludzi, 2 konta Premium.
1 miesiąc za darmo. 24,99 zł/mies.

<https://www.spotify.com/>

Spotify

Spotify is all the music you'll ever need.

Download Spotify

Spotify is a digital music service that gives you access to millions ...

Sign up

Sign up with your email address ·
By clicking on sign-up, you ...

More results from [spotify.com](https://www.spotify.com/)

Premium

Get one month free. Ad-free music listening. Get over 50 million ...

Free Mobile App

Spotify is a digital music service that gives you access to millions ...

People also ask

Is Spotify really free?



How can I use Spotify for free?



Is Spotify music illegal?



<https://twitter.com/Spotify>

Spotify (@Spotify) · Twitter

Drop your comfort album 📌
#WorldMentalHealthDay

Twitter · 19 hours ago



It's okay to take a break 📌
We hope this World Mental Health Day, you take a beat and tune into you.

Twitter · 2 days ago



@aespa_official are always leveling up and their first Mini Album #Savage is proof 📌
spotify.link/savage

Twitter · 3 days ago

Top stories

Sydney Morning Herald

IVF on us: Spotify leads the way with family benefits for Australian staff

1 day ago



The Indian Express

Spotify Premium gets Diwali discount: Here are the details

8 hours ago



View all

Systemy rekomendacyjne:

- wybierają podzbiór dostępnych elementów
- optymalizują zadaną miarę
 - zysk
 - zadowolenie użytkowników
- automatyzują proces doboru treści na stronach
- zastępują człowieka
 - nie męczą się
 - analizują większą ilość danych
 - są bezstronne (w pewnym sensie)

Co możemy optymalizować?

- Miary syntetyczne:

- accuracy,
- precision,
- recall,
- F1,
- ...

- Miary biznesowe:

- CTR (click-through ratio),
- czas spędzony na czytaniu/słuchaniu/oglądaniu,
- głębokość scrolla,
- przychody z reklam,
- liczba subskrypcji,
- powracalność użytkowników,
- długość i liczba sesji,
- ...

Jakimi danymi dysponujemy?

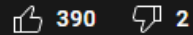
Dane od użytkowników (feedback)

- Feedback bezpośredni:

- Thumb-up



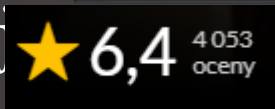
- Polubienie



- Reakcja



- Oceny



- Udostępnienia

- Komentarze

- Feedback pośredni:

- Impresje

- Kliknięcia

- Brak kliknięć

- Czas spędzony

- Głębokość scrolla

- Powracalność (np. do kolejnych odcinków)

Dane dotyczące treści

- Dane o autorach, czasie utworzenia, miejscu pochodzenia
- Analiza tekstu (podobieństwo tekstów, named entities, sentiment)
- Tagi (gatunek filmu, muzyki, tematyka artykułu, rodzaj piwa)
- Parametry (taneczność lub instrumentalność muzyki, mroczność lub komediowość filmu)

Algorytmy naiwne

Losowy

- Losowa próbka ze zbioru elementów rekomendowanych
- Należy rozważyć, czy prawdopodobieństwo wylosowania każdego elementu powinno być takie samo
- Warto wyeliminować niektóre elementy (np. Już przeczytane/obejrzane)

Naj-

- Najczęściej czytany, najlepiej oceniany, najbardziej kontrowersyjny, ...
- Warto rozważyć zawężenie grupy docelowej (np. "Najpopularniejsze wśród fanów Black Mirror")
- Prosty i efektywny
- Zaniedbuje nowe materiały

Algorytm inż. Mamonia

- Najbardziej podobne do tych, które użytkownik chętnie czytał/oglądał/słuchał
- W jaki sposób możemy porównywać elementy między sobą?
- W krótkiej perspektywie może dać świetne rezultaty
- Użytkownik w końcu się znudzi

Content-based

Definicja problemu

- Zbiór użytkowników $\{1, 2, \dots, n_u\}$
- Zbiór filmów $\{1, 2, \dots, n_m\}$
- Użytkownicy oceniają filmy w skali (na przykład) $\{1, 2, 3, 4, 5\}$
- $r(i, j) = 1$ - jeśli użytkownik j ocenił film i
 - $r(i, j) = 0$ - w p. p.
- $y^{(i, j)}$ - ocena filmu i wystawiona przez użytkownika j
 - Wartość zdefiniowana tylko wtedy, gdy $r(i, j) == 1$

Definicja problemu

- Każdy film i opisuje wektor parametrów $X^i = (1, x_1^i, \dots x_k^i)$
- Nie wnioskujemy, skąd te parametry się wzięły

Film	Alice	Bob	Carol	Dave	x_1 (romantyczność)	x_2 (akcja)
Listy do M.	5	5	0	0	0.9	0.2
Notting Hill	5	?	?	0	1.0	0.0
Wesele	0	0	5	4	0.3	0.7
Szklana Pułapka	0	0	5	?	0.05	1.0

Definicja problemu

- Chcemy przewidzieć, jakie będą wartości brakujących ocen dla każdego z użytkowników
- Dla każdego użytkownika chcemy wyznaczyć wektor preferencji $\Theta^j = (\theta^0_j, \theta^1_j, \dots, \theta^k_j)$ tak, by dla każdego filmu i i użytkownika j , iloczyn $(\theta^j)^T X^i$ jak najlepiej przewidywał wystawioną ocenę

Film	Alice	Bob	Carol	Dave	x_1 (romantyczność)	x_2 (akcja)
Listy do M.	5	5	0	0	0.9	0.2
Notting Hill	5	?	?	0	1.0	0.0
Wesele	0	0	5	4	0.3	0.7
Szklana Pułapka	0	0	5	?	0.05	1.0

Jak znaleźć Θ^j

- Załóżmy, że w jakiś sposób (np. losowo) wysnaczyliśmy wartość Θ^j
- Możemy policzyć, jak bardzo pomyliliśmy się w wycenie filmu i w porównaniu z rzeczywistą oceną:
$$\text{err} = (\Theta^j)^T X^i - y^{(i, j)}$$
- Naszym celem jest minimalizacja tych błędów dla każdej rzeczywistej oceny filmu i i wystawionej przez użytkownika j
- Dodatkowo, chcielibyśmy otrzymać możliwie małe wartości współrzędnych Θ^j

Optimization objective:

To learn $\theta^{(j)}$ (parameter for user j):

$$\min_{\theta^{(j)}} \frac{1}{2} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{k=1}^n (\theta_k^{(j)})^2$$

To learn $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$:

$$\min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

Optimization algorithm:

$$\min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

Gradient descent update:

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \sum_{i:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) x_k^{(i)} \quad (\text{for } k = 0)$$

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \left(\sum_{i:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) x_k^{(i)} + \lambda \theta_k^{(j)} \right) \quad (\text{for } k \neq 0)$$

Jak ocenić jakość
rekomendacji?

Dlaczego to nie jest takie proste?

- Standardowa procedura - podział na zbiór treningowy i testowy, porównanie predykcji wygenerowanych przez model ze zbiorem testowym
- Skąd mamy wiedzieć, jak bardzo nasz system się myli?
- Bezwzględne wartości nic nam nie mówią, musimy je porównać z jakimiś punktami odniesienia (ang. *baseline*)
- Dobry punkt odniesienia stanowią naiwne algorytmy (zwłaszcza losowy jest powszechnie używany, ale trzeba to robić z głową)

Rekomendacje pozytywne i negatywne

- Podczas oceny wytrenowanego modelu, obliczamy liczbę rekomendacji prawdziwie dodatnich (TP), fałszywie dodatnich (FP), prawdziwie ujemnych (TN) i fałszywie ujemnych (FN)
- Na nasze potrzeby możemy uznać, że wyniki pozytywne to np. oceny 4* i 5*, a negatywne to oceny 1* i 2*

Miary syntetyczne

- $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$
- $\text{Precision} = (\text{TP}) / (\text{TP} + \text{FP})$
- $\text{Recall} = (\text{TP}) / (\text{TP} + \text{FN})$
- $\text{F1} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

Czy to aby nie przypadek?

- Pojedynczy pomiar jest podatny na przypadkowe błędy
- Warto przeprowadzić wiele pomiarów, za każdym razem dzieląc dane na zbiór treningowy i testowy za każdym razem w inny sposób
- Jeśli w większości przypadków wartość miary, którą chcemy porównywać (np. recall) jest wyższa dla wytrenowanego modelu niż dla punktu odniesienia - możemy uznać, że omdel jest lepszy

A ma pan dowód?

- Wnioskowanie "na oko" nie zawsze jest wystarczające, a co ważniejsze - nie przystoi przyszłym inżynierom
- Aby mieć pewność*, że różnica badanej miary jest istotna statystycznie - możemy przeprowadzić test statystyczny

Test znaku

- Najprostszy test statystyczny do porównywania dwóch zmiennych
- Hipoteza zerowa – wytrenowany model jest tak samo dobry jak punkt odniesienia
- Hipoteza alternatywna – wytrenowany model jest lepszy od punktu odniesienia
- Obieramy poziom istotności statystycznej p (np. $p = 0.05$)

Test znaku

- Wykonujemy n (niezależnych!) prób
- Dla każdej próby obliczamy, czy różnica między wynikiem wytrenowanego modelu a wynikiem punktu odniesienia jest dodatnia czy ujemna

Nr próby	Recall(model)	Recall(baseline)	Różnica
1	0.72	0.51	+
2	0.54	0.56	-
...
n	0.66	0.47	+

Test znaku

- Załóżmy, że na $n=20$ prób, w 17 nasz model okazał się lepszy
- Jakie jest prawdopodobieństwo takiego zdarzenia (co najmniej 17 pozytywnych prób), gdy prawdziwa jest hipoteza zerowa, czyli że model jest równie dobry co punkt odniesienia?
 - Obliczamy to tak samo, jak prawdopodobieństwo wyrzucenia orła
 - 17 dodatnich w 20 próbach: ~ 0.00109
 - 18 dodatnich w 20 próbach: ~ 0.00018
 - 19 dodatnich w 20 próbach: ~ 0.00002
 - 20 dodatnich w 20 próbach: ~ 0.00000
- Prawdopodobieństwo, że przy dwóch porównywalnych algorytmach jeden z nich okaże się lepszy od drugiego w 17 próbach na 20 wynosi około 0.00129
- Prawdopodobieństwo to jest mniejsze od $p = 0.05$, zatem możemy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną - nasz algorytm jest istotnie lepszy

Podsumowanie

- Czym są systemy rekomendacyjne?
- Jakie dane możemy zbierać?
- Jakie miary możemy optymalizować?
- Content-based recommender
- Jak porównać jakość algorytmów?