

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Wprowadzenie

dr inż. Aleksandra Karpus

4 października 2023



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESA RADY MINISTRÓW
MINISTERSTWA ROZWOJU ECONOMICZNEGO

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

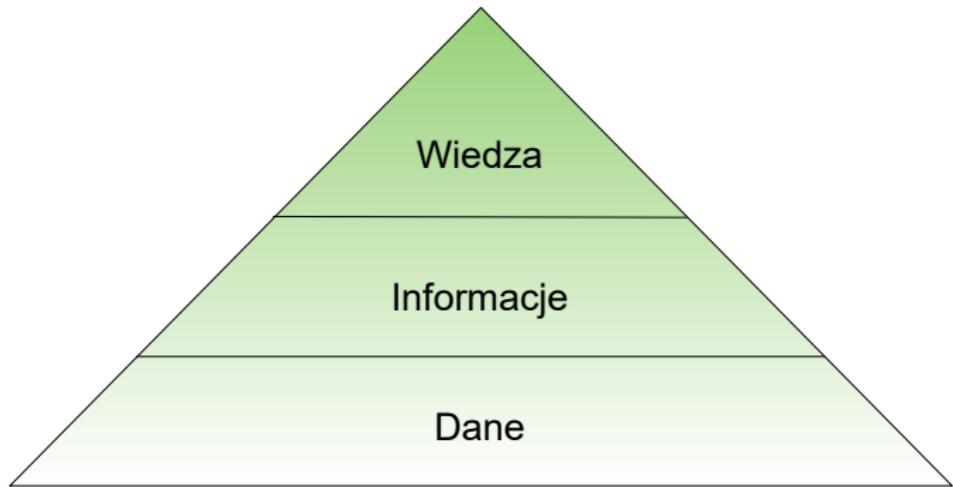
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Organizacja zajęć

- 30 godz. wykładu, 15 godz. laboratorium, 15 godz. projektu
- Zajęcia laboratoryjne?
- Zajęcia projektowe odbywają się na przemian*

- Wszystkie części należy zaliczyć na min. 51%
- 2 kolokwia zaliczające wykład: w połowie i pod koniec semestru
- Udział w ocenie:
 - wykład - 40%
 - laboratorium - 30%
 - projekt - 30%
- Sposób obliczania oceny:
 - 51-60% – 3.0
 - 61-70% – 3.5
 - 71-80% – 4.0
 - 81-90% – 4.5
 - 91-100% – 5.0

Czym jest "wiedza"?

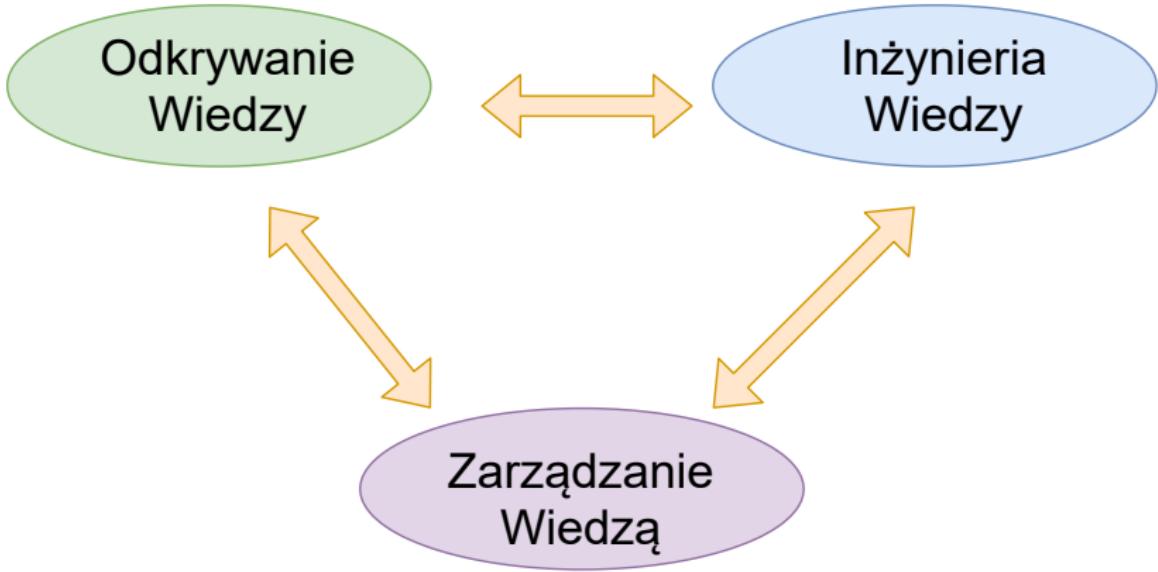


- "Fakty zgromadzone z obserwacji lub zapisów dotyczących zjawisk, obiektów lub ludzi." (Clare and Loucopoulos, 1987)
- "Dane reprezentują nieustrukturyzowane fakty." (Avison and Fitzgerald, 1995)

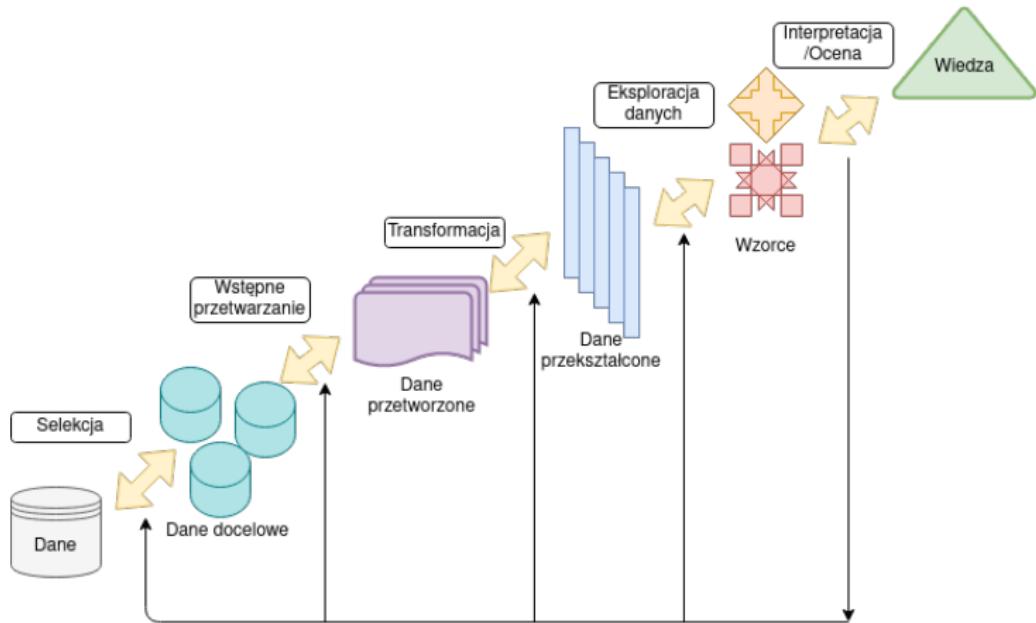
- "Informacje to to, co powstaje w wyniku pewnych działań myślowych człowieka (obserwacji, analiz) z sukcesem zastosowanych do danych by odkryć ich istotę lub znaczenie." (Galland, 1982)
- "Informacja ma znaczenie... pochodzi z wyselekcjonowania danych, ich podsumowania i prezentacji w taki sposób, by były użyteczne dla odbiorcy." (Avison and Fitzgerald, 1995)
- "Dane, które zostały ukształtowane lub uformowane przez człowieka w istotną i użyteczną postać" (Laudon and Laudon, 1991)

- "Wiedza ludzka jest informacją połączoną z doświadczeniem, kontekstem, interpretacją i refleksją." (Davenport and Prusak, 1998)
- "Wiedza komputerowa jest to zbiór informacji zapisanych w pamięci komputera wraz ze zdolnością komputera do samodzielnego poszerzania tego zbioru drogą wnioskowania." (Goczyła, 2011)

Czym jest "odkrywanie wiedzy"?



Proces odkrywania wiedzy



- ETL
- Hurtownie Danych
- Bazy Wiedzy
 - Ontologie
 - Logika Opisowa
 - Wnioskowanie
- Eksploracja Danych
- Pozyskiwanie Informacji

Czym jest?

Eksploracja i analiza dużej ilości danych w celu odkrycia nieznanych lub ukrytych, ale zrozumiałych informacji oraz użycie ich do podejmowania decyzji biznesowych i ich wdrażania poprzez formułowanie taktycznych i strategicznych inicjatyw oraz ocenę ich sukcesu.^a

^aDefinicja zaczerpnięta z wykładów do przedmiotu Eksploracja Danych autorstwa dra W. Waloszka.

Proces ekploracji danych



By Kenneth Jensen - Own work based on:
<ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf> (Figure 1), CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=24930610>

Eksploracyjna analiza danych

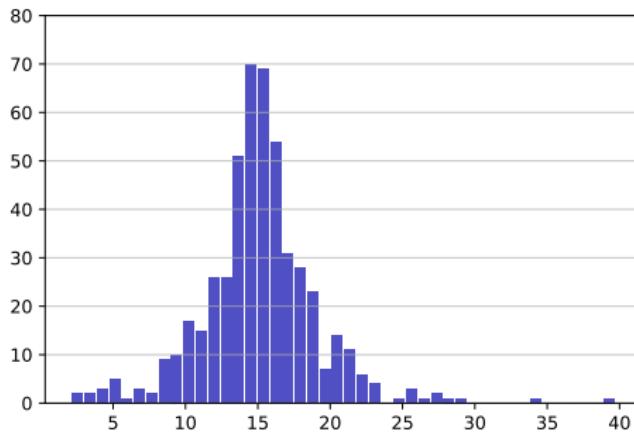
- zestaw interaktywnych i wizualnych technik analizy zbioru danych bez bardziej precyzyjnego wskazania celu eksploracji;
- bardzo ważny element etapu wstępnej analizy danych
- pozwala na:
 - zgłębienie zbioru danych,
 - sprawdzenie zależności między atrybutami,
 - identyfikację nietypowych przykładów bądź podzbiorów przykładów,
 - opracowanie wstępnych wniosków dotyczących prawidłowości w zbiorze przykładów.

Główne narzędzia wizualizacji

- histogramy,
- wykresy pudełkowe (ang. *box plot*),
- wykresy punktowe (ang. *scatter plot*),
- mapy konturowe (ang. *contour plot*),
- macierze wykresów punktowych.

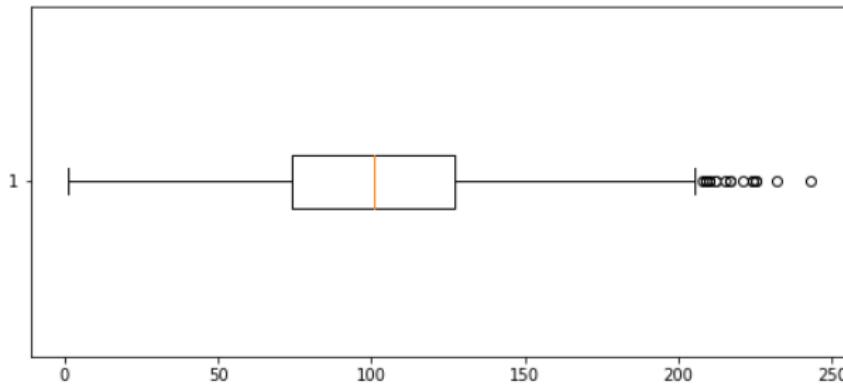
Histogram

- graficzny sposób przedstawiania rozkładu empirycznego cechy;
- pozwala na:
 - przybliżenie kształtu rozkładu,
 - identyfikację punktów oddalonych,
 - identyfikację wartości specjalnych.



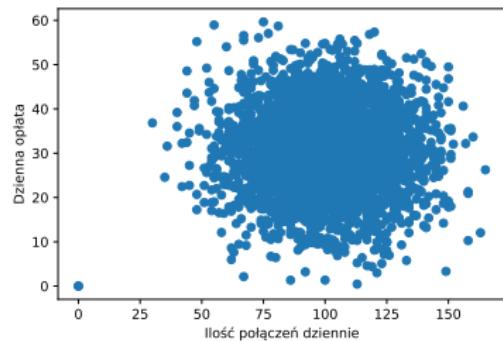
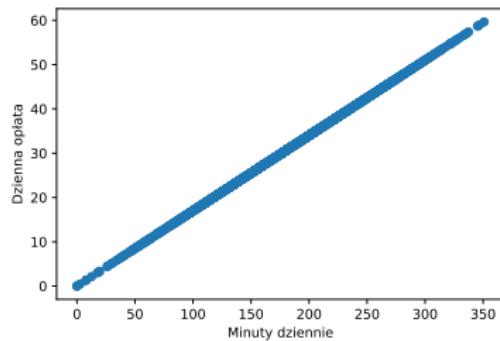
Wykres pudełkowy

- służy do przedstawienia podstawowych parametrów rozkładu wartości atrybutów numerycznych;
- pudełko reprezentuje pierwszy, drugi i trzeci kwartyl (Q_1 , Q_2 i Q_3);
- wąsy pokazują rozrzuć danych „poza pudełkiem” - najczęściej jest to najbardziej skrajna wartość mieszcząca się jeszcze w odległości $1,5 * (Q_3 - Q_1)$ od odpowiednio Q_1 i Q_3 ;
- pozostałe punkty oznaczone są jako oddalone.



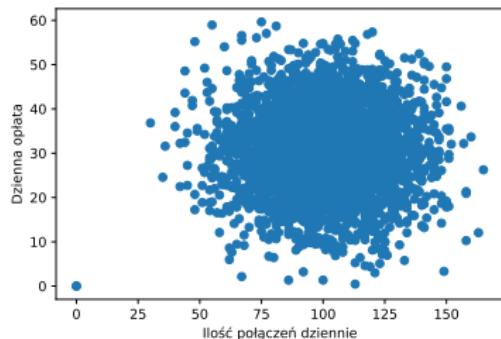
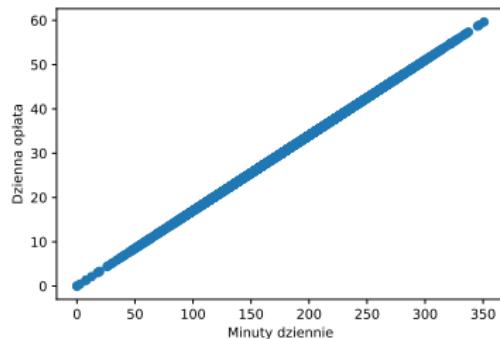
Wykresy punktowe

- pozwalają na wizualizację zależności pomiędzy parą zmiennych;
- pozwala na:
 - odkrywanie korelacji między parą zmiennych,
 - identyfikację punktów oddalonych od ogólnego trendu.



Wykresy punktowe

- pozwalają na wizualizację zależności pomiędzy parą zmiennych;
- pozwala na:
 - odkrywanie korelacji między parą zmiennych,
 - identyfikację punktów oddalonych od ogólnego trendu.



Macierze wykresów punktowych
pozwalają na badanie większych podzbiorów atrybutów.

Przygotowanie danych

Przygotowanie danych

- uzupełnienie/usuwanie wartości brakujących,
- analizę punktów oddalonych,
- normalizację wartości atrybutów,
- wyrównanie reprezentacji wszystkich interesujących klas przykładów,
- korektę typu atrybutów na potrzeby odpowiedniego modelu,
- zmniejszanie liczby wymiarów.

Dlaczego?

Dlaczego?

- Wartości brakujące mogą wskazywać na niską jakość przykładu.
- Brak wartości oznaczany wartością specjalną (np. 0) może wprowadzać zaburzenia w dalszej analizie danych.
- Część metod tworzenia modeli nie obsługuje dobrze wartości brakujących lub obsługuje je w specyficzny sposób.

Dlaczego?

- Wartości brakujące mogą wskazywać na niską jakość przykładu.
- Brak wartości oznaczany wartością specjalną (np. 0) może wprowadzać zaburzenia w dalszej analizie danych.
- Część metod tworzenia modeli nie obsługuje dobrze wartości brakujących lub obsługuje je w specyficzny sposób.

Jak?

Dlaczego?

- Wartości brakujące mogą wskazywać na niską jakość przykładu.
- Brak wartości oznaczany wartością specjalną (np. 0) może wprowadzać zaburzenia w dalszej analizie danych.
- Część metod tworzenia modeli nie obsługuje dobrze wartości brakujących lub obsługuje je w specyficzny sposób.

Jak?

- uzupełnianie wartością średnią, medianą, wartością najczęściej występującą itp.,
- uzupełnianie wartością losową,
- usunięcie przykładu.

Dlaczego?

Dlaczego?

- Punkty oddalone mogą stanowić dane błędne.
- Poprawne dane opisywane przez punkty oddalone często mogą mieć nieoczekiwany wpływ na tworzenie modelu.

Dlaczego?

- Punkty oddalone mogą stanowić dane błędne.
- Poprawne dane opisywane przez punkty oddalone często mogą mieć nieoczekiwany wpływ na tworzenie modelu.

Jak?

Dlaczego?

- Punkty oddalone mogą stanowić dane błędne.
- Poprawne dane opisywane przez punkty oddalone często mogą mieć nieoczekiwany wpływ na tworzenie modelu.

Jak?

- usunięcie przykładów stanowiących punkty oddalone,
- wydzielenie punktów oddalonych jako osobne zbiory przykładów do dalszej analizy.

Dlaczego?

Dlaczego?

- Zróżnicowane przedziały wartości dla różnych atrybutów mogą mieć negatywny wpływ na zastosowanie niektórych metod.
- Są metody, które wymagają normalizacji.

Dlaczego?

- Zróżnicowane przedziały wartości dla różnych atrybutów mogą mieć negatywny wpływ na zastosowanie niektórych metod.
- Są metody, które wymagają normalizacji.

Jak?

Dlaczego?

- Zróżnicowane przedziały wartości dla różnych atrybutów mogą mieć negatywny wpływ na zastosowanie niektórych metod.
- Są metody, które wymagają normalizacji.

Jak?

- normalizacja min-max,
- standaryzacja.

Dlaczego?

Dlaczego?

Znaczne dysproporcje w liczbie poszczególnych klas wyznaczanych przez atrybut celu mogą sprawić, że najbardziej interesujące klienta klasy zostaną zignorowane przez model.

Dlaczego?

Znaczne dysproporcje w liczbie poszczególnych klas wyznaczanych przez atrybut celu mogą sprawić, że najbardziej interesujące klienta klasy zostaną zignorowane przez model.

Jak?

Dlaczego?

Znaczne dysproporcje w liczbie poszczególnych klas wyznaczanych przez atrybut celu mogą sprawić, że najbardziej interesujące klienta klasy zostaną zignorowane przez model.

Jak?

- usunięcie części przykładów nadreprezentowanych klas,
- generacja pewnej liczby przykładów klas niedostatecznie reprezentowanych,
- wydzielenie osobnych klas do osobnej analizy.

Zmiana typu atrybutów

Dlaczego?

Zmiana typu atrybutów

Dlaczego?

Istnieją metody, które wymagają użycia konkretnego rodzaju atrybutu.

Zmiana typu atrybutów

Dlaczego?

Istnieją metody, które wymagają użycia konkretnego rodzaju atrybutu.

Jak?

Dlaczego?

Istnieją metody, które wymagają użycia konkretnego rodzaju atrybutu.

Jak?

- dyskretyzacja wartości liczbowych,
- nadanie wartości numerycznych (zachowanie porządku dla atrybutów porządkowych).

Zadania eksploracji danych

Zadania eksploracji danych

- Predykcja

- Predykcja
 - Klasyfikacja

- Predykcja
 - Klasyfikacja
 - Regresja

Zadania eksploracji danych

- Predykcja
 - Klasyfikacja
 - Regresja
- Deskrypcja

Zadania eksploracji danych

- Predykcja
 - Klasyfikacja
 - Regresja
- Deskrypcja
- Odkrywanie zależności

- Predykcja
 - Klasyfikacja
 - Regresja
- Deskrypcja
- Odkrywanie zależności
- Analiza szeregów czasowych i trendów

Bibliografia

1. K. Goczyła, *Ontologie w systemach informatycznych*, EXIT, Warszawa, 2011.
2. W. Waloszek i T. Zawadzka, *Eksploracja danych*, materiały do przedmiotu, Politechnika Gdańska, 2020.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Predykcja i ewaluacja modelu

dr inż. Aleksandra Karpus

alekarpu@pg.edu.pl
pokój 631 EA

4 października 2023



Preidykcja

Preidykcja

- regresja

Preidykcja

- regresja
- klasyfikacja

Preidykcja

- regresja - służy do przewidywania atrybutów numerycznych (ciągłych),
- klasyfikacja

Preidykcja

- regresja - służy do przewidywania atrybutów numerycznych (ciągłych),
- klasyfikacja - służy do przewidywania atrybutów nominalnych (klas).

Regresja liniowa

Regresja liniowa

dopasowanie danych do funkcji liniowej (przekształcenia aficznego):

$$y = a * X + b ,$$

gdzie:

y - zmienna objaśniana,

X - zmienna/zmienne objaśniające,

a - współczynnik kierunkowy,

b - wyraz wolny.

Regresja liniowa

dopasowanie danych do funkcji liniowej (przekształcenia aficznego):

$$y = a * X + b ,$$

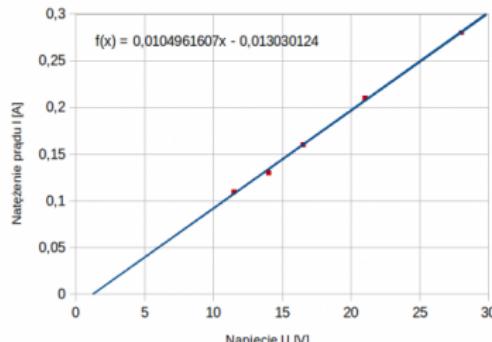
gdzie:

y - zmienna objaśniana,

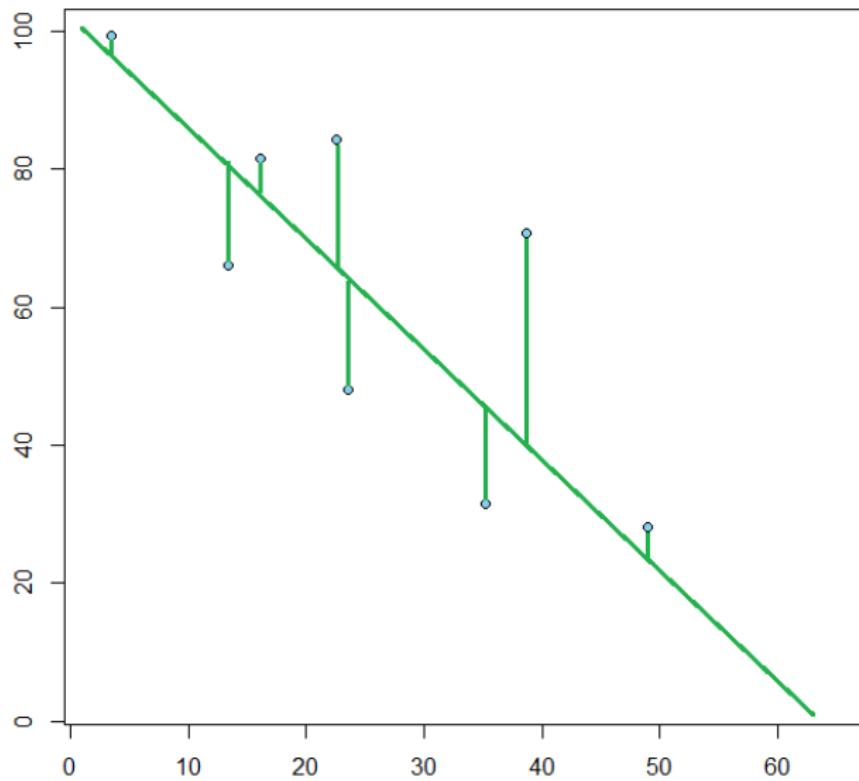
X - zmienna/zmienne objaśniające,

a - współczynnik kierunkowy,

b - wyraz wolny.



Ocena modelu



Ocena modelu - miary

Ocena modelu - miary

- błąd średniokwadratowy

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - x_i)^2 ,$$

Ocena modelu - miary

- błąd średniokwadratowy

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - x_i)^2 ,$$

- średni błąd bezwzględny

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{x}_i - x_i| ,$$

Ocena modelu - miary

- błąd średniokwadratowy

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - x_i)^2 ,$$

- średni błąd bezwzględny

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{x}_i - x_i| ,$$

- współczynnik determinacji

$$R^2 = \frac{\sum_{i=1}^n (\hat{x}_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} .$$

Ocena modelu - problemy

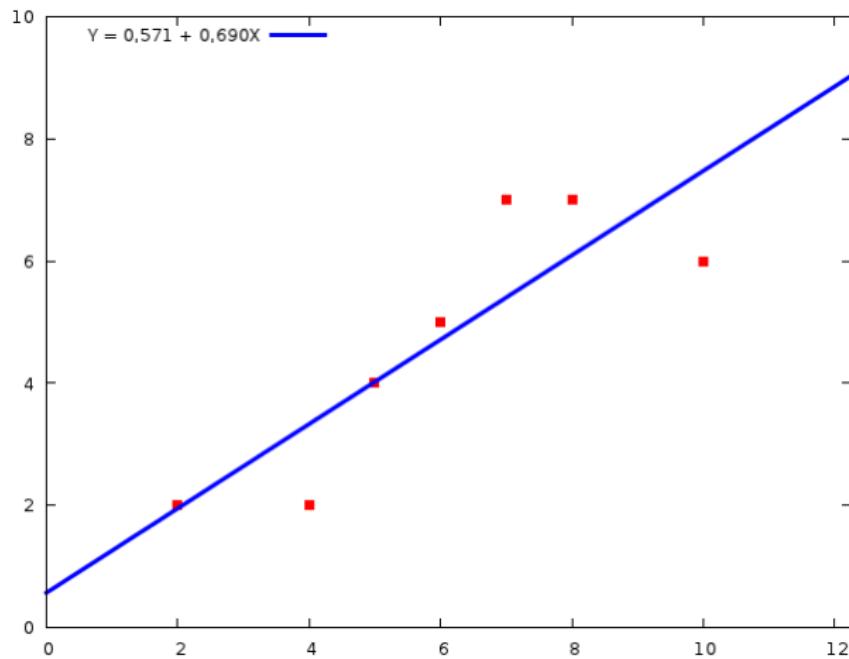
Ocena modelu - problemy

- nadmierne dopasowanie do danych (ang. *overfitting*),

Ocena modelu - problemy

- nadmierne dopasowanie do danych (ang. *overfitting*),
- niedostateczne dopasowanie do danych (ang. *underfitting*).

Underfitting



Underfitting - rozwiązanie

Underfitting - rozwiązanie

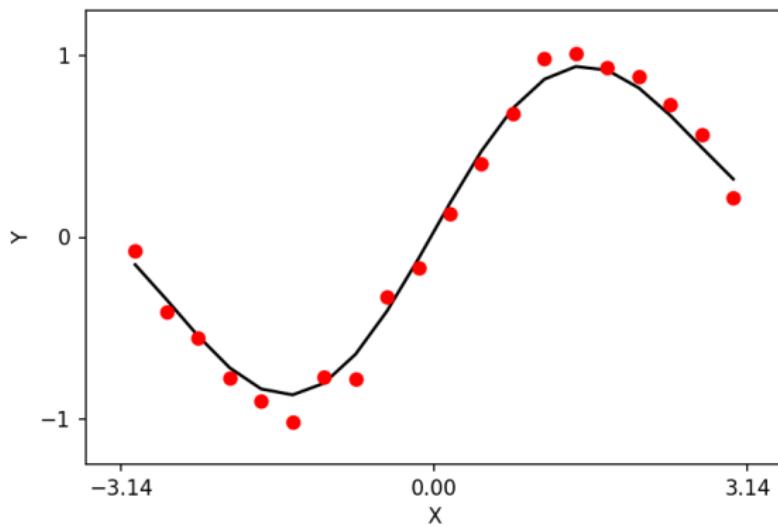
Wprowadzenie wielomianowości

poprzez dodanie nowych zmiennych opisujących $x_n = f_n(x)$, np.
 $f_n(x) = x^n$.

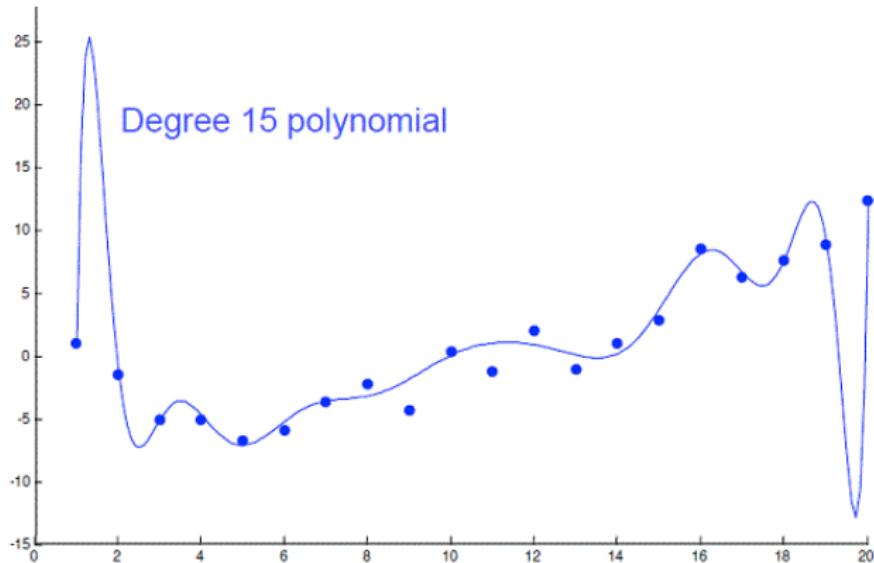
Underfitting - rozwiązanie

Wprowadzenie wielomianowości

poprzez dodanie nowych zmiennych opisujących $x_n = f_n(x)$, np.
 $f_n(x) = x^n$.



Overfitting



Overfitting - rozwiązania

Overfitting - rozwiązania

- regularyzacja,

Overfitting - rozwiązania

- regularyzacja,
- przygotowanie danych do oceny modelu:

Overfitting - rozwiązania

- regularyzacja,
- przygotowanie danych do oceny modelu:
 - podział zbioru na uczący i testowy (ang. *hold-out validation*),

Overfitting - rozwiązania

- regularyzacja,
- przygotowanie danych do oceny modelu:
 - podział zbioru na uczący i testowy (ang. *hold-out validation*),
 - walidacja skrośna (ang. *cross-validation*),

Overfitting - rozwiązania

- regularyzacja,
- przygotowanie danych do oceny modelu:
 - podział zbioru na uczący i testowy (ang. *hold-out validation*),
 - walidacja skrośna (ang. *cross-validation*),
 - *leave-one-out*,

Overfitting - rozwiązania

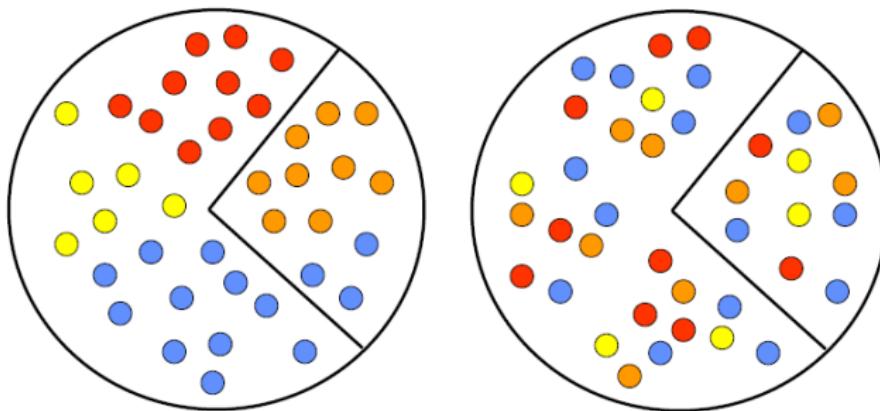
- regularyzacja,
- przygotowanie danych do oceny modelu:
 - podział zbioru na uczący i testowy (ang. *hold-out validation*),
 - walidacja skrośna (ang. *cross-validation*),
 - *leave-one-out*,
 - 0.632 *bootstrap*.

Hold-out validation

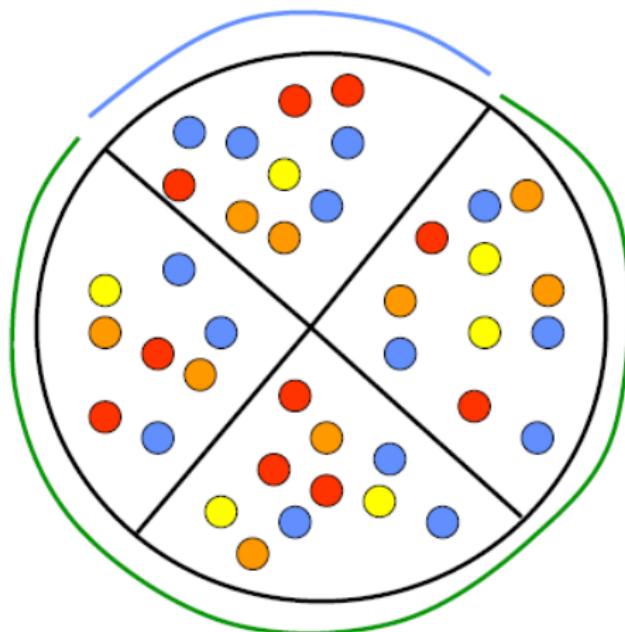
- podział losowy,
- podział ze stratyfikacją.

Hold-out validation

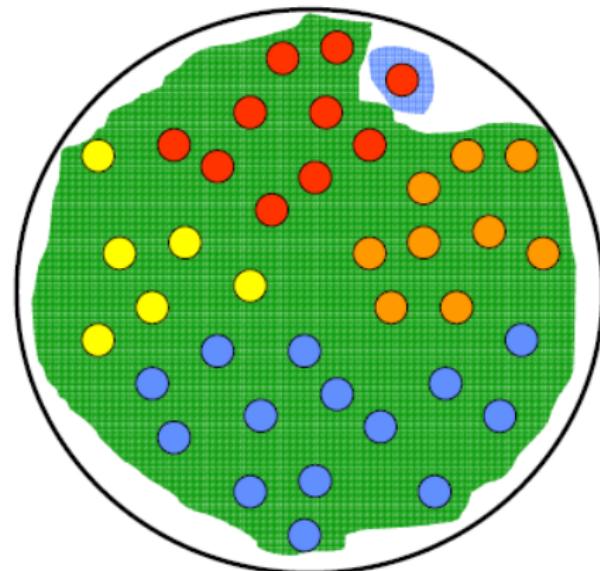
- podział losowy,
- podział ze stratyfikacją.



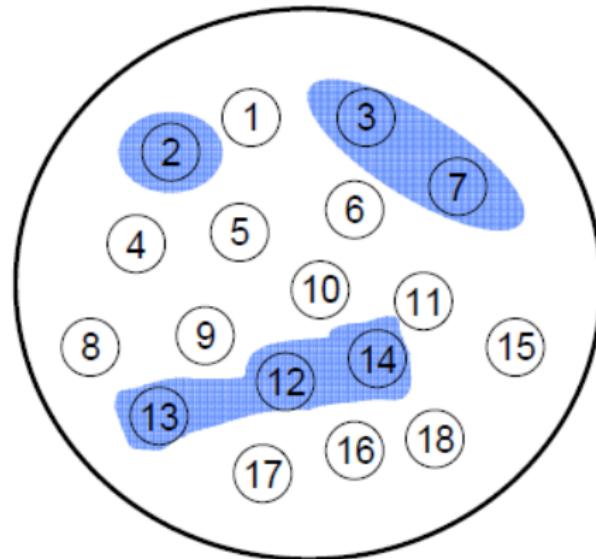
Walidacja skrośna



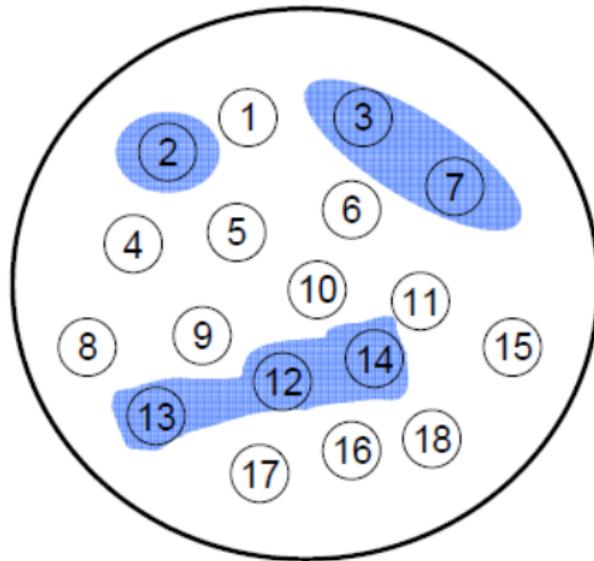
Leave-one-out



0.632 bootstrap



0.632 bootstrap



Liczba przykładów niewybranych stanowi statystycznie $0.368 k$ (wybranych 0.632).

Regularyzacja

Regularyzacja

- grzbietowa (L_2),

Regularyzacja

- grzbietowa (L_2),
- Lasso (L_1).

Regularizacja Lasso (L1)

- W przypadku zwykłej regresji liniowej szukane są wagi minimalizujące funkcję celu postaci

$$f(w) = \sum_{i=1}^n (y_i - w^T x_i)^2$$

Regularizacja Lasso (L1)

- W przypadku zwykłej regresji liniowej szukane są wagi minimalizujące funkcję celu postaci

$$f(w) = \sum_{i=1}^n (y - w^T x_i)^2$$

- Dla regresji Lasso dodany jest składnik (funkcja kary) ograniczający wartości wag

$$f(w) = \sum_{i=1}^n (y - w^T x_i)^2 + \lambda |w_i|$$

Regularizacja Lasso (L1)

- W przypadku zwykłej regresji liniowej szukane są wagi minimalizujące funkcję celu postaci

$$f(w) = \sum_{i=1}^n (y - w^T x_i)^2$$

- Dla regresji Lasso dodany jest składnik (funkcja kary) ograniczający wartości wag

$$f(w) = \sum_{i=1}^n (y - w^T x_i)^2 + \lambda |w_i|$$

- Składnikiem regularyzującym jest norma L1, czyli suma wartości bezwzględnych wag.

Regularizacja grzbietowa (Ridge lub L2)

- Dla regresji grzbietowej składnikiem regularyzującym jest norma L2

$$f(w) = \sum_{i=1}^n (y - w^T x_i)^2 + \lambda ||w||_2$$

Regularizacja grzbietowa (Ridge lub L2)

- Dla regresji grzbietowej składnikiem regularyzującym jest norma L2

$$f(w) = \sum_{i=1}^n (y - w^T x_i)^2 + \lambda ||w||_2$$

- Norma L2 to suma kwadratów wartości wag

$$||w||_2 = \sqrt{\sum_{i=1}^n w_i^2}$$

Metoda wektorów nośnych (SVM)

Metoda wektorów nośnych (SVM)

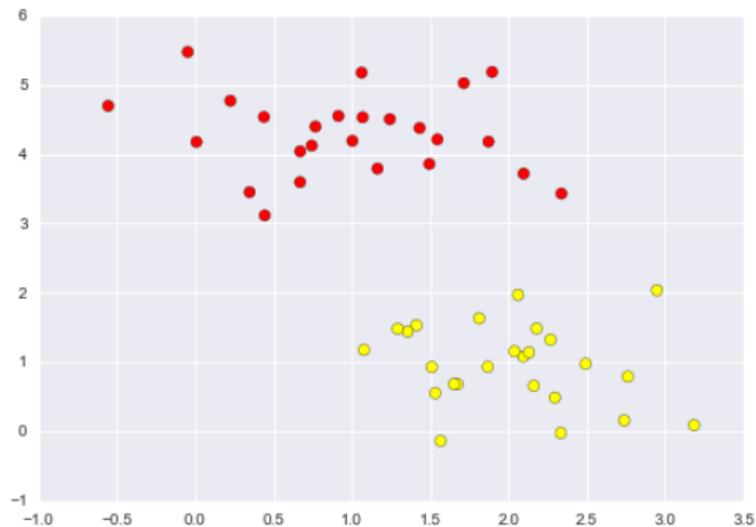
- ang. Support Vector Machine,

Metoda wektorów nośnych (SVM)

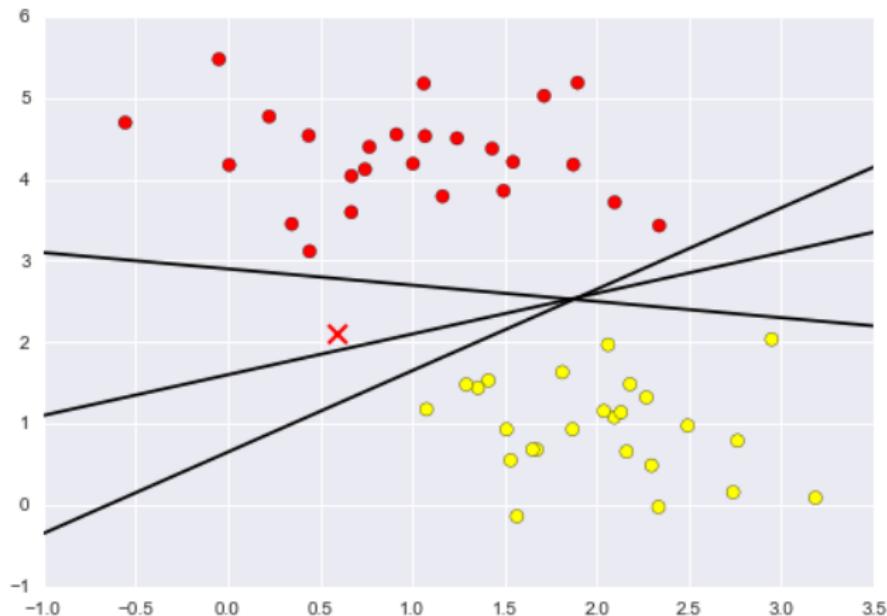
- ang. Support Vector Machine,
- służy do klasyfikacji.

Metoda wektorów nośnych (SVM)

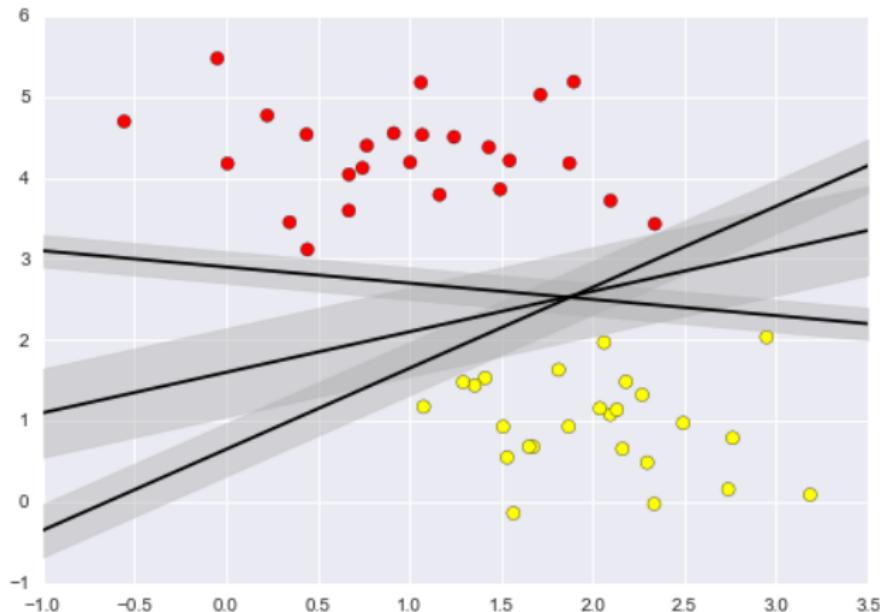
- ang. Support Vector Machine,
- służy do klasyfikacji.



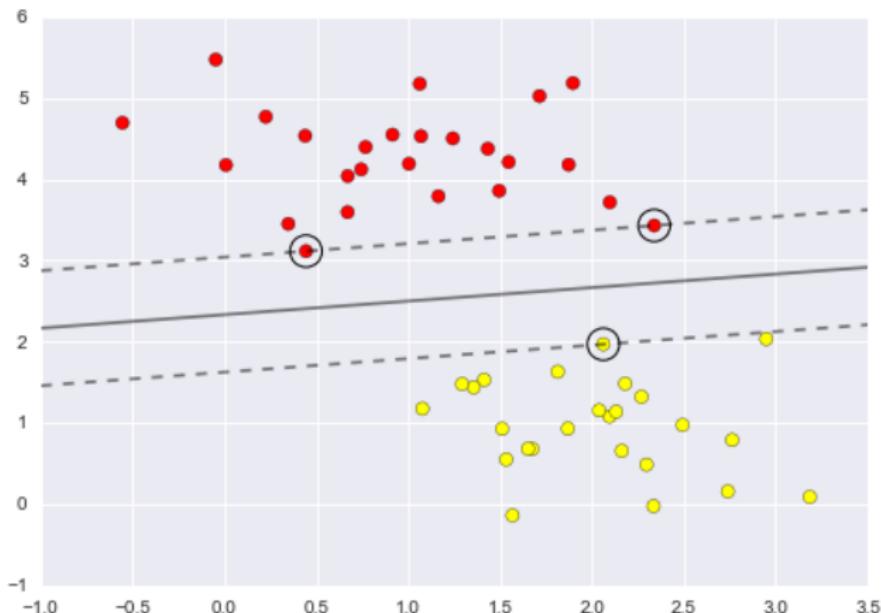
Metoda wektorów nośnych (SVM)



Metoda wektorów nośnych (SVM)



Metoda wektorów nośnych (SVM)



Dziękuję za uwagę!

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Wprowadzenie do systemów rekomendacyjnych

dr inż. Aleksandra Karpus

11 października 2023



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESIA RADY MINISTRÓW

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)».

Agenda

1. Idea systemów rekomendacyjnych
2. Zarys historyczny
3. Rodzaje systemów rekomendacyjnych
4. Macierz ocen

Dlaczego warto zajmować się systemami rekomendacji?

Co się dzieje przez minutę w Internecie?



Czym są systemy rekomendacyjne?

System rekomendacyjny

system filtrowania informacji wspierający użytkownika w danej sytuacji decyzyjnej poprzez zawężenie zestawu możliwych opcji i priorytetyzację jego elementów. Priorytetyzacja może opierać się na jawnie lub niejawnie wyrażonych preferencjach użytkownika, a także na wcześniejszych zachowaniach użytkowników o podobnych preferencjach.

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]
- 1994 - Fab [4] - pierwszy hybrydowy algorytm

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]
- 1994 - Fab [4] - pierwszy hybrydowy algorytm
- 1998 - Google - powstanie algorytmu PageRank

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]
- 1994 - Fab [4] - pierwszy hybrydowy algorytm
- 1998 - Google - powstanie algorytmu PageRank
- 1998 - Amazon [5] - item-based collaborative filtering

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]
- 1994 - Fab [4] - pierwszy hybrydowy algorytm
- 1998 - Google - powstanie algorytmu PageRank
- 1998 - Amazon [5] - item-based collaborative filtering
- 2006 - wyzwanie Netflix¹

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]
- 1994 - Fab [4] - pierwszy hybrydowy algorytm
- 1998 - Google - powstanie algorytmu PageRank
- 1998 - Amazon [5] - item-based collaborative filtering
- 2006 - wyzwanie Netflix¹
- 2009 - 1 mln dolarów - nagroda Netflixa

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]
- 1994 - Fab [4] - pierwszy hybrydowy algorytm
- 1998 - Google - powstanie algorytmu PageRank
- 1998 - Amazon [5] - item-based collaborative filtering
- 2006 - wyzwanie Netflix¹
- 2009 - 1 mln dolarów - nagroda Netflix
- 2010 - rekomendacje na YouTube

¹<https://www.netflixprize.com/>

Kamienie milowe systemów rekomendacji:

- 1979 - Grundy [1] - bibliotekarz
- 1990 - Tapestry [2] - "collaborative filtering"
- 1992 - GroupLens [3]
- 1994 - Fab [4] - pierwszy hybrydowy algorytm
- 1998 - Google - powstanie algorytmu PageRank
- 1998 - Amazon [5] - item-based collaborative filtering
- 2006 - wyzwanie Netflix¹
- 2009 - 1 mln dolarów - nagroda Netflix
- 2010 - rekomendacje na YouTube
- 2017 - Neural Collaborative Filtering [6]

¹<https://www.netflixprize.com/>

Rodzaje systemów rekomendacyjnych

- Niespersonalizowane

- Niespersonalizowane
- Spersonalizowane

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)
 - User-based

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)
 - User-based
 - Item-based

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)
 - User-based
 - Item-based
 - Rozkład macierzy

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (CARS)

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Oparte na *Uczeniu Głębokim* (DLRS)

- Niespersonalizowane
- Spersonalizowane
 - Content-based filtering (CBF)
 - Collaborative filtering (CF)
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Oparte na *Uczeniu Głębokim* (DLRS)
 - Inne

Oceny użytkowników

| Użytkownik | P1 | P2 | P3 | P4 | P5 |
|------------|----|----|----|----|----|
| Alicja | 3 | 1 | 4 | 4 | 3 |
| Bogdan | 4 | 2 | 0 | 4 | 5 |
| Cezary | 1 | 5 | 5 | 4 | 3 |
| Daniel | 5 | 3 | 4 | 0 | 4 |
| Ewa | 3 | 0 | 2 | 1 | 2 |

Macierz ocen użytkownika

| | | | | |
|---|---|---|---|---|
| 3 | 1 | 4 | 4 | 3 |
| 4 | 2 | 0 | 4 | 5 |
| 1 | 5 | 5 | 4 | 3 |
| 5 | 3 | 4 | 0 | 4 |
| 3 | 0 | 2 | 1 | 2 |

Macierz R

| | | | | | |
|----------|--|---|----------|--|---|
| | | | <i>i</i> | | |
| | | 0 | | | 0 |
| <i>u</i> | | | r_{ui} | | |
| | | | | | 0 |
| | | 0 | | | |

Rekomendacja jako funkcja

$$R : \text{Użytkownicy} \times \text{Produkty} \mapsto \text{Oceny}$$

Problemy zapisu macierzowego

- gęstość macierzy < 0.01

Problemy zapisu macierzowego

- gęstość macierzy < 0.01
 - Netflix ≈ 0.002

- gęstość macierzy < 0.01
 - Netflix ≈ 0.002
- stronniczość użytkowników (ang. user bias)

- gęstość macierzy < 0.01
 - Netflix ≈ 0.002
- stronniczość użytkowników (ang. user bias)
- dobór skali ocen

- gęstość macierzy < 0.01
 - Netflix ≈ 0.002
- stronniczość użytkowników (ang. user bias)
- dobór skali ocen
 - przedział

- gęstość macierzy < 0.01
 - Netflix ≈ 0.002
- stronniczość użytkowników (ang. user bias)
- dobór skali ocen
 - przedział
 - parzysta czy nie?

Bibliografia

1. E. Rich (1979): User modeling via stereotypes, *Cognitive Science*, Vol. 3, No. 4, pp. 329–354.
2. D. Goldberg, B. Oki, D. Nichols, D. B. Terry (1992): Using Collaborative Filtering to Weave an Information Tapestry, *Communications of the ACM*, December, Vol. 35, No. 12, pp. 61-70.
3. P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, J. Riedl (1994): GroupLens: An open architecture for collaborative filtering of netnews, In *Proceedings of the ACM Conf. Computer Support Cooperative Work (CSC)*, pp. 175-186.
4. M. Balabanovic, Y. Shoham (1997): Fab: Content-based, Collaborative Recommendation, *Communications of the ACM*, Vol.40, No.3, pp.66-72.
5. B. Smith, G. Linden (2017): Two Decades of Recommender Systems at Amazon.com, in *IEEE Internet Computing*, vol. 21, no. 3, pp. 12-18, doi: 10.1109/MIC.2017.72.
6. X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.S. Chua (2017): Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 173–182. DOI:<https://doi.org/10.1145/3038912.3052569>

1. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2010. *Recommender Systems Handbook* (1st ed.). Springer-Verlag New York, Inc., New York, NY, USA.
2. Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction* (1st ed.). Cambridge University Press, New York, NY, USA.
3. Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer International Publishing.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

k najbliższych sąsiadów

dr inż. Aleksandra Karpus

11 i 18 października 2023



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESA RADY MINISTRÓW

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)».

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{11} & r_{12} & \cdots & r_{1m} \\ \cdots & \cdots & \ddots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}$$

- n – liczba użytkowników, m – liczba produktów,
- r_{ui} – ocena wystawiona produktowi i przez użytkownika u
- ocena może być dana explicite, np. na skali pięciogwiazdkowej (pozytywna i negatywna) lub implicite (tylko pozytywna).

- **Niespersonalizowane**

- Spersonalizowane

- Filtrowanie oparte na zawartości (CBF)
- Filtrowanie kolaboracyjne (CF)
 - User-based
 - Item-based
 - Rozkład macierzy
- Kontekstowe (CARS)
- Wielodziedzinowe (CDRS)
- Oparte na *Uczeniu Głębokim* (DLRS)
- Inne

Niespersonalizowane algorytmy rekomendacyjne

Niespersonalizowane algorytmy rekомендacyjne

- Najpopularniejsze

Niespersonalizowane algorytmy rekомендacyjne

- Najpopularniejsze
- Najwyżej oceniane

Najpopularniejsze produkty

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 4 | 4 | 3 |
| 4 | 0 | 0 | 4 | 5 |
| 1 | 0 | 5 | 0 | 3 |
| 0 | 3 | 4 | 0 | 4 |
| 3 | 0 | 2 | 1 | 2 |

| | | | | |
|---|--|--|--|--|
| 3 | | | | |
|---|--|--|--|--|

Najpopularniejsze produkty

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 4 | 4 | 3 |
| 4 | 0 | 0 | 4 | 5 |
| 1 | 0 | 5 | 0 | 3 |
| 0 | 3 | 4 | 0 | 4 |
| 3 | 0 | 2 | 1 | 2 |

| | | | | |
|---|---|---|---|---|
| 3 | 1 | 4 | 3 | 5 |
|---|---|---|---|---|

Najpopularniejsze produkty

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 4 | 4 | 3 |
| 4 | 0 | 0 | 4 | 5 |
| 1 | 0 | 5 | 0 | 3 |
| 0 | 3 | 4 | 0 | 4 |
| 3 | 0 | 2 | 1 | 2 |

| | | | | | |
|---|---|---|---|----------|--|
| 3 | 1 | 4 | 3 | 5 | |
|---|---|---|---|----------|--|

Najwyżej oceniane produkty

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 5 | 4 | 4 |
| 4 | 0 | 0 | 4 | 5 |
| 2 | 0 | 5 | 0 | 3 |
| 0 | 5 | 4 | 0 | 5 |
| 3 | 0 | 2 | 1 | 3 |

$$n_i = \frac{\sum_{U_i} r_{ui}}{|U_i|}$$

| | | | | |
|---|--|--|--|--|
| 3 | | | | |
|---|--|--|--|--|

Najwyżej oceniane produkty

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 5 | 4 | 4 |
| 4 | 0 | 0 | 4 | 5 |
| 2 | 0 | 5 | 0 | 3 |
| 0 | 5 | 4 | 0 | 5 |
| 3 | 0 | 2 | 1 | 3 |

$$n_i = \frac{\sum_{U_i} r_{ui}}{|U_i|}$$

| | | | | |
|---|----------|---|---|---|
| 3 | 5 | 4 | 3 | 4 |
|---|----------|---|---|---|

Najwyżej oceniane z pokryciem

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 5 | 4 | 4 |
| 4 | 0 | 0 | 4 | 5 |
| 2 | 0 | 5 | 0 | 3 |
| 0 | 5 | 4 | 0 | 5 |
| 3 | 0 | 2 | 1 | 3 |

$$n_i = \frac{\sum_{U_i} r_{ui}}{|U_i|+C}$$

$$C = 1$$

$$n_1 = \frac{4+2+3}{3+1}$$

| | | | | |
|------|--|--|--|--|
| 2,25 | | | | |
|------|--|--|--|--|

Najwyżej oceniane z pokryciem

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 5 | 4 | 4 |
| 4 | 0 | 0 | 4 | 5 |
| 2 | 0 | 5 | 0 | 3 |
| 0 | 5 | 4 | 0 | 5 |
| 3 | 0 | 2 | 1 | 3 |

$$n_i = \frac{\sum_{U_i} r_{ui}}{|U_i|+C}$$

$$C = 1$$

$$n_1 = \frac{5}{1+1}$$

| | | | | |
|------|-----|--|--|--|
| 2,25 | 2,5 | | | |
|------|-----|--|--|--|

Najwyżej oceniane z pokryciem

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 5 | 4 | 4 |
| 4 | 0 | 0 | 4 | 5 |
| 2 | 0 | 5 | 0 | 3 |
| 0 | 5 | 4 | 0 | 5 |
| 3 | 0 | 2 | 1 | 3 |

$$n_i = \frac{\sum_{U_i} r_{ui}}{|U_i|+C}$$
$$C = 1$$

| | | | | |
|------|-----|-----|------|--------------|
| 2,25 | 2,5 | 3,2 | 2,25 | 3,(6) |
|------|-----|-----|------|--------------|

- Niespersonalizowane
- Spersonalizowane
 - Filtrowanie oparte na zawartości (CBF)
 - **Filtrowanie kolaboracyjne (CF)**
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Oparte na *Uczeniu Głębokim* (DLRS)
 - Inne

Filtrowanie kolaboracyjne użytkownik-użytkownik [6]

- Bazuje na założeniu, że użytkownicy o podobnej historii ocen będą lubili te same rzeczy w przyszłości.

Filtrowanie kolaboracyjne użytkownik-użytkownik [6]

- Bazuje na założeniu, że użytkownicy o podobnej historii ocen będą lubili te same rzeczy w przeszłości.
- Punkt wyjścia: wersja niespersonalizowana - najwyżej oceniane:

$$S(u, i) = \frac{\sum_{v \in U} r_{vi}}{|U|}$$

Wersja spersonalizowana

$$S(u, i) = \frac{\sum_{v \in U} r_{vi} * w_{uv}}{\sum_{v \in U} w_{uv}}$$

A co ze stronniczością użytkowników?

Radzenie sobie ze stronniczością użytkowników

Wersja niespersonalizowana

$$S(u, i) = \bar{r}_u + \frac{\sum_{v \in U} (r_{vi} - \bar{r}_v)}{|U|}$$

Radzenie sobie ze stronniczością użytkowników

Wersja niespersonalizowana

$$S(u, i) = \bar{r}_u + \frac{\sum_{v \in U} (r_{vi} - \bar{r}_v)}{|U|}$$

Wersja spersonalizowana

$$S(u, i) = \bar{r}_u + \frac{\sum_{v \in U} (r_{vi} - \bar{r}_v) w_{uv}}{\sum_{v \in U} w_{uv}}$$

Czym jest waga w_{uv} ?

Podobieństwo użytkowników

- Zwykle mierzone poprzez korelację lub dystans.
- Najpopularniejsze:
 - Współczynnik korelacji Pearsona

$$sim(u, v) = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}} , \quad (1)$$

gdzie i oznacza produkt ze zbioru I , $r_{u,i}$ jest oceną przedmiotu i daną przez użytkownika u , zaś \bar{r}_u oznacza średnią ocenę użytkownika u .

- Podobieństwo cosinusowe

$$cos(u, v) = \frac{u \bullet v}{\|u\| \|v\|} , \quad (2)$$

gdzie \bullet oznacza iloczyn skalarny, zaś $\|u\|$ jest normą wektora u .

- Współczynnik korelacji rang Spearmana.

Czy chcemy brać pod uwagę wszystkich użytkowników?

Dlaczego nie?

Wady uwzględniania wszystkich użytkowników

- koszt obliczeniowy,
- negatywny wpływ ocen użytkowników o sprzecznych gustach/poglądach,
- nie ma sensu wykonywać obliczeń dla użytkownika $v = u$.

dowolny podzbiór zbioru użytkowników wybrany poprzez arbitralnie ustalone kryterium.

dowolny podzbiór zbioru użytkowników wybrany poprzez arbitralnie ustalone kryterium.

Kryteria wyboru sąsiedztwa:

- Wybór losowy,
- Ustalona liczba sąsiadów,
- Minimalne podobieństwo,
- Wspólny klaster.

Mamy dany zbiór produktów I , zbiór użytkowników U i rzadką macierz ocen R . Aby wyznaczyć ocenę $s(u, i)$ dla użytkownika u i produktu i , należy wykonać poniższe kroki.

- Dla każdego użytkownika $v \neq u$ wyznaczyć w_{uv} .
- Wybrać sąsiedztwo $V \in U$ (np. z najwyższą wartością w_{uv}).
- Obliczyć predykcję:

$$s(u, i) = \bar{r}_u + \frac{\sum_{v \in V} (r_{vi} - \bar{r}_v) w_{uv}}{\sum_{v \in V} w_{uv}}$$

Problemy podejścia użytkownik-użytkownik

- Rzadkość macierzy ocen
- Koszt obliczeniowy wyznaczania korelacji pomiędzy użytkownikami
- Potrzeba aktualizacji preferencji użytkownika w czasie rzeczywistym

- Niespersonalizowane
- Spersonalizowane
 - Filtrowanie oparte na zawartości (CBF)
 - **Filtrowanie kolaboracyjne (CF)**
 - User-based
 - **Item-based**
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Oparte na *Uczeniu Głębokim* (DLRS)
 - Inne

Dlaczego filtrowanie kolaboracyjne bazujące na produkcie ma szansę działać lepiej od tego bazującego na użytkowniku?

- Zwykle mamy ograniczoną liczbę produktów i nieskończenie wielu użytkowników.
- Produkty mają wyższą średnią ilość ocen niż użytkownicy.
- Odświeżanie w czasie rzeczywistym korelacji pomiędzy produktami nie jest tak kluczowe jak w przypadku użytkowników.

Algorytm (Item kNN)

Mamy dany zbiór produktów I , zbiór użytkowników U i rzadką macierz ocen R . Aby wyznaczyć ocenę $s(u, i)$ dla użytkownika u i produktu i , należy wykonać poniższe kroki.

- Dla każdego produktu $j \neq i$ wyznaczyć w_{ij} .
- Wybrać sąsiedztwo $J \in I$.
- Obliczyć predykcję:

$$s(u, i) = \frac{\sum_{j \in J} r_{uj} * w_{ij}}{\sum_{j \in J} w_{ij}}$$

- Co rozumiemy poprzez podobieństwo produktów?

- Co rozumiemy poprzez podobieństwo produktów?
- W jaki sposób wybrać sąsiedztwo J ?

- Co rozumiemy poprzez podobieństwo produktów?
- W jaki sposób wybrać sąsiedztwo J ?
- Czy ocena użytkownika wymaga normalizacji, tak jak w przypadku User k NN?

Co rozumiemy poprzez podobieństwo produktów?

- Podobieństwo cosinusowe na znormalizowanych wektorach
- Współczynnik korelacji Pearsona

- Przede wszystkim muszą to być produkty ocenione przez użytkownika.
- Można ograniczyć do k najlepszych poprzez wartość podobieństwa.

Algorytm (Item kNN) po normalizacji

Mamy dany zbiór produktów I , zbiór użytkowników U i rzadką macierz ocen R . Aby wyznaczyć ocenę $s(u, i)$ dla użytkownika u i produktu i , należy wykonać poniższe kroki.

- Dla każdego produktu $j \neq i$ wyznaczyć w_{ij} .
- Wybrać sąsiedztwo $J \in I$.
- Obliczyć predykcję:

$$s(u, i) = \bar{r}_i + \frac{\sum_{j \in J} (r_{uj} - \bar{r}_j) w_{ij}}{\sum_{j \in J} w_{ij}}$$

Czy coś się zmieni dla oceny danej nie wprost (np. kliknięcie, zakup)?

Item k NN dla oceny nie wprost

- Podobieństwo cosinusowe lub prawdopodobieństwo warunkowe.
- Brak normalizacji - ewentualna normalizacja do wektorów jednostkowych.
- Element r_{uj} nic nie wnosi.

$$s(u, i) = \sum_{j \in J} w_{ij}$$

- Rozwiążanie nie sprawdzi się w przypadku produktów sezonowych, np. ozdoby świąteczne.
- W przypadku, w którym mamy więcej produktów niż użytkowników w systemie.
- Zakłada się tutaj stabilne preferencje użytkowników.
- Wyniki mają mniejsze *serendipity* niż te zwrócone przez rozwiązania bazujące na użytkowniku.

Jakie są różnice pomiędzy algorytmem k najbliższych sąsiadów bazującym na użytkowniku i bazującym na produkcie? Jakie są ich wady i zalety?

Bibliografia

1. J. A. Konstan, M. D. Ekstrand. *Introduction to Recommender Systems*. University of Minnesota. Coursera.
2. P. Cremonesi. *Basic Recommender Systems*. EIT Digital & Politecnico di Milano. Coursera.
3. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2010. *Recommender Systems Handbook* (1st ed.). Springer-Verlag New York, Inc., New York, NY, USA.
4. Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction* (1st ed.). Cambridge University Press, New York, NY, USA.
5. Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer International Publishing.
6. J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. 1999. An algorithmic framework for performing collaborative filtering. In *Proc. of the 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR '99, page 230–237, New York, NY, USA.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Ewaluacja systemów rekomendacyjnych

dr inż. Aleksandra Karpus

25 października 2023

Agenda

1. Idea oceny jakości systemów rekomendacyjnych
2. Rodzaje miar jakości
3. Miary jakości
4. Protokół oceny systemu rekomendacyjnego

Podejścia do ewaluacji systemów rekomendacyjnych

- Retrospektywne - z wykorzystaniem danych historycznych
- Prospektywne - z udziałem użytkowników

Ogólna idea ewaluacji retrospektywnej systemów rekomendacyjnych

- Wykorzystanie istniejących danych.
- Zasymulowanie użytkownika.
- Sprawdzenie, czy system jest w stanie przewidzieć zachowanie użytkownika.

Ewaluacja systemów rekomendacyjnych a zadania rekomendacyjne

- Predykcja oceny użytkownika - spojrzenie lokalne
- Tworzenie listy rekomendacji (ranking) - spojrzenie porównawcze

Co możemy mierzyć?

- Jak bardzo przewidywana ocena różni się od tej, którą dał użytkownik?
 - Potrzeba agregacji wyników.
- Czy wybrane przez użytkownika produkty znalazły się na liście rekomendowanej przez system?
- Na której pozycji się znalazły?

- Miary dokładności predykcji
- Miary z dziedziny pozyskiwania informacji
- Miary dedykowane liście najlepszych rekomendacji
- Inne miary

- Miary błędu

- Miary błędu
- Podejście "*pomiń jeden*" (ang. *leave-one-out*)

- Miary błędu
- Podejście "*pomiń jeden*" (ang. *leave-one-out*)
- Nie możemy ocenić skuteczności predykcji dla produktów, których użytkownik nie zna.

- Średni błąd bezwzględny - **MAE**

$$\text{MAE} = \frac{1}{|T|} \sum_{(u,i) \in T} |\hat{r}_{u,i} - r_{u,i}| , \quad (1)$$

gdzie T - zbiór testowy, u i i - użytkownik i produkt ze zbioru testowego,

$\hat{r}_{u,i}$ - ocena wyznaczona przez algorytm, a $r_{u,i}$ - prawdziwa ocena użytkownika.

- Średni błąd bezwzględny - **MAE**

$$\text{MAE} = \frac{1}{|T|} \sum_{(u,i) \in T} |\hat{r}_{u,i} - r_{u,i}| , \quad (1)$$

gdzie T - zbiór testowy, u i i - użytkownik i produkt ze zbioru testowego,

$\hat{r}_{u,i}$ - ocena wyznaczona przez algorytm, a $r_{u,i}$ - prawdziwa ocena użytkownika.

- Średni błąd kwadratowy - **MSE**

$$\text{MSE} = \frac{1}{|T|} \sum_{(u,i) \in T} (\hat{r}_{u,i} - r_{u,i})^2 . \quad (2)$$

- Średni błąd bezwzględny - **MAE**

$$\text{MAE} = \frac{1}{|T|} \sum_{(u,i) \in T} |\hat{r}_{u,i} - r_{u,i}| , \quad (1)$$

gdzie T - zbiór testowy, u i i - użytkownik i produkt ze zbioru testowego,

$\hat{r}_{u,i}$ - ocena wyznaczona przez algorytm, a $r_{u,i}$ - prawdziwa ocena użytkownika.

- Średni błąd kwadratowy - **MSE**

$$\text{MSE} = \frac{1}{|T|} \sum_{(u,i) \in T} (\hat{r}_{u,i} - r_{u,i})^2 . \quad (2)$$

- Pierwiastek z błędu średniokwadratowego - **RMSE**

$$\text{RMSE} = \sqrt{\frac{1}{|T|} \sum_{(u,i) \in T} (\hat{r}_{u,i} - r_{u,i})^2} . \quad (3)$$

Przykład

| Produkt | Ocena |
|---------|-------|
| i1 | 2 |
| i2 | 4 |
| i3 | 5 |
| i4 | 3 |

Tabela: Przykładowy zbiór testowy.

| Produkty | SR1 | SR2 |
|----------|-----|-----|
| i1 | 4 | 2 |
| i2 | 2 | 4 |
| i3 | 3 | 1 |
| i4 | 3 | 3 |

Tabela: Predykcje zwrócone przez dwa systemy rekomendacyjne, SR1 i SR2 dla zbioru testowego z powyższej tabeli.

- Błędy dla SR1:

$$\text{MAE} = \frac{1}{4}(|4 - 2| + |2 - 4| + |3 - 5| + |3 - 3|) = 1.5 , \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{4}(|4 - 2|^2 + |2 - 4|^2 + |3 - 5|^2 + |3 - 3|^2)} \approx 1.7 . \quad (5)$$

- Błędy dla SR1:

$$\text{MAE} = \frac{1}{4}(|4 - 2| + |2 - 4| + |3 - 5| + |3 - 3|) = 1.5 , \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{4}(|4 - 2|^2 + |2 - 4|^2 + |3 - 5|^2 + |3 - 3|^2)} \approx 1.7 . \quad (5)$$

- Błędy dla SR2:

$$\text{MAE} = \frac{1}{4}(|2 - 2| + |4 - 4| + |1 - 5| + |3 - 3|) = 1 , \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{4}(|2 - 2|^2 + |4 - 4|^2 + |1 - 5|^2 + |3 - 3|^2)} = 2 . \quad (7)$$

- Agregacja po:
 - ocenach - wszystko razem,
 - użytkownikach,
 - produktach,
 - innych kryteriach (ale o tym później).

- Agregacja po:
 - ocenach - wszystko razem,
 - użytkownikach,
 - produktach,
 - innych kryteriach (ale o tym później).
- Miary błędu to zły wybór w przypadkach, w których interesuje nas co użytkownik lubi, a nie to, czego i w jakim stopniu nie lubi, np. tworzenie kolejki utworów muzycznych.

Przykład 2

Rekomendacje w formie listy z predykcjami

- Zbiór filmów ocenionych w skali 5-cio gwiazdkowej.
- Film X , który użytkownik u oceniłby na 2,5 gwiazdek, gdyby go widział (ale nie widział).
- System rekommendacji filmów, który zwraca listę filmów z przewidywaną oceną i myli się średnio o 1,5 gwiazki.
- Czy użytkownik u jest zadowolony z podjętej decyzji o obejrzeniu/nieobejrzeniu filmu X ?

- Precyza

$$\text{Precyza} = \frac{|\{\text{istotne produkty}\} \cap \{\text{zarekomendowane produkty}\}|}{|\{\text{zarekomendowane produkty}\}|}, \quad (8)$$

- Precyza

$$\text{Precyza} = \frac{|\{\text{istotne produkty}\} \cap \{\text{zarekomendowane produkty}\}|}{|\{\text{zarekomendowane produkty}\}|}, \quad (8)$$

- Czułość (ang. *sensitivity/recall*)

$$\text{Czułość} = \frac{|\{\text{istotne produkty}\} \cap \{\text{zarekomendowane produkty}\}|}{|\{\text{istotne produkty}\}|}. \quad (9)$$

- Precyza

$$\text{Precyza} = \frac{|\{\text{istotne produkty}\} \cap \{\text{zarekomendowane produkty}\}|}{|\{\text{zarekomendowane produkty}\}|}, \quad (8)$$

- Czułość (ang. *sensitivity/recall*)

$$\text{Czułość} = \frac{|\{\text{istotne produkty}\} \cap \{\text{zarekomendowane produkty}\}|}{|\{\text{istotne produkty}\}|}. \quad (9)$$

- Miara F1 - średnia harmoniczna powyższych

$$F1 = 2 \cdot \frac{\text{Precyza} \cdot \text{Czułość}}{\text{Precyza} + \text{Czułość}}. \quad (10)$$

Przykład 3

Lista top-5 rekomendacji dla Alicji:

| Film | Ocena |
|--------------|-------|
| Spiderman | 2 |
| Donnie Darko | 4 |
| Incepcja | 5 |
| Turysta | 3 |
| Zielona Mila | 4 |

Tabela: Zbiór testowy dla Alicji.

1. Przerwana lekcja muzyki
2. *Donnie Darko*
3. *Doktor Parnassus*
4. *Adwokat*
5. *Incepcja*
6. *Turysta*
7. *Zielona Mila*
8. *Armagedon*
9. *Shrek*
10. *Spiderman*

Przykład 3 (c.d.)

$$\text{Precyzja} = \frac{2}{5}$$

$$\text{Czułość} = \frac{2}{3}$$

$$F1 = \frac{1}{2}$$

Macierz błędów

| | | Przewidywane | |
|-----------|---------------------------|---------------------------|--|
| Faktyczne | Lubi | Nie lubi | |
| Lubi | Prawdziwie pozytywne (TP) | Fałszywie negatywne (FN) | |
| Nie lubi | Fałszywie pozytywne (FP) | Prawdziwie negatywne (TN) | |

Macierz błędów

| | | Przewidywane | |
|-----------|---------------------------|---------------------------|--|
| Faktyczne | Lubi | Nie lubi | |
| Lubi | Prawdziwie pozytywne (TP) | Fałszywie negatywne (FN) | |
| Nie lubi | Fałszywie pozytywne (FP) | Prawdziwie negatywne (TN) | |

-

$$\text{Precyza} = \frac{TP}{TP + FP}$$

| | | Przewidywane | |
|-----------|---------------------------|---------------------------|--|
| Faktyczne | Lubi | Nie lubi | |
| Lubi | Prawdziwie pozytywne (TP) | Fałszywie negatywne (FN) | |
| Nie lubi | Fałszywie pozytywne (FP) | Prawdziwie negatywne (TN) | |

-

$$\text{Precyzja} = \frac{TP}{TP + FP}$$

- Czułość, czyli współczynnik wyników prawdziwie pozytywnych (TPR)

$$TPR = \frac{TP}{TP + FN}$$

Macierz błędów

| | | Przewidywane | |
|-----------|---------------------------|---------------------------|--|
| Faktyczne | Lubi | Nie lubi | |
| Lubi | Prawdziwie pozytywne (TP) | Fałszywie negatywne (FN) | |
| Nie lubi | Fałszywie pozytywne (FP) | Prawdziwie negatywne (TN) | |

-

$$\text{Precyzja} = \frac{TP}{TP + FP}$$

- Czułość, czyli współczynnik wyników prawdziwie pozytywnych (TPR)

$$TPR = \frac{TP}{TP + FN}$$

- Współczynnik wyników fałszywie pozytywnych (FPR)

$$FPR = \frac{FP}{FP + TN}$$

Precyza - podsumowanie

$$\text{Precyza} = \frac{|\{\text{istotne produkty}\} \cap \{\text{zarekomendowane produkty}\}|}{|\{\text{zarekomendowane produkty}\}|}$$

$$\text{Precyza} = \frac{TP}{TP + FP}$$

Przykład 3 (c.d.)

$$\text{Precyzja} = \frac{TP}{TP + FP} = \frac{2}{2 + 3} = \frac{2}{5}$$

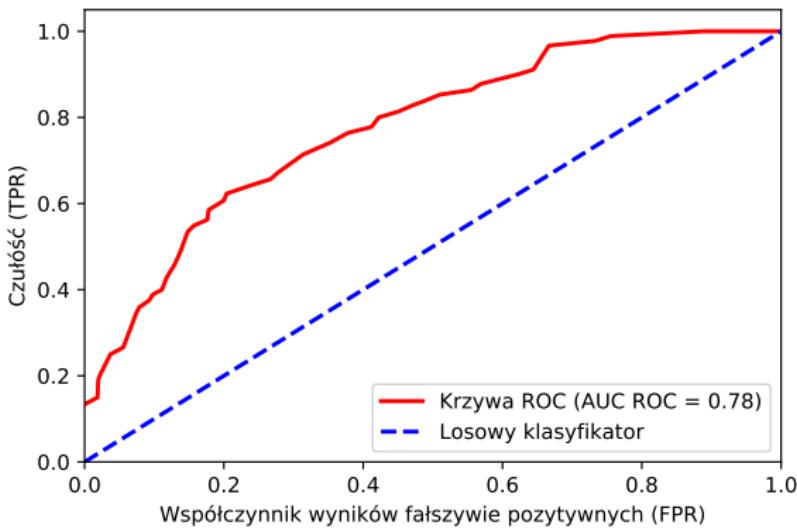
$$\text{Czułość} = \frac{TP}{TP + FN} = \frac{2}{2 + 1} = \frac{2}{3}$$

$$\text{FPR} = \frac{FP}{FP + TN} = \frac{3}{3 + 2} = \frac{3}{5}$$

Miary z dziedziny pozyskiwania informacji

ROC AUC

Charakterystyka operacyjna odbiornika (ang. *receiver operating characteristic*) lub krzywa ROC



Miary dedykowane liście najlepszych rekomendacji

Średnia odwrotność rankingu (ang. *mean reciprocal rank*, MRR)

$$\text{MRR} = \frac{1}{N} \sum_{l=1}^N \frac{1}{\text{rank}_l} , \quad (11)$$

gdzie N to liczba list rekomendacji, dla których liczymy średnią, rank_l odpowiada pozycji pierwszego istotnego produktu na l -tej liście rekomendacji.

Miary dedykowane liście najlepszych rekomendacji

Średnia precyzja (ang. *average precision*, AP)[4]

$$AP = \frac{\sum_{k=1}^n (\text{precyzja}@k \cdot \text{istotny}(k))}{|\{\text{istotne produkty}\}|}, \quad (12)$$

gdzie n jest rozmiarem listy rekomendacji, a $\text{istotny}(k)$ jest funkcją, która zwraca 1 jeżeli produkt na pozycji k na liście jest istotny i zero w przeciwnym przypadku .

MAP (ang. Mean Average Precision)

$$MAP = \frac{\sum_{l=1}^N AP(l)}{N}, \quad (13)$$

gdzie N to liczba list rekomendacji.

Miary dedykowane liście najlepszych rekommendacji

DCG (ang. *Discounted Cumulative Gain*)

$$\text{DCG} = \frac{1}{N} \sum_{u=1}^N \sum_{k=1}^n \frac{g_{ui_k}}{\max(1, \log_b k)} , \quad (14)$$

gdzie N oznacza liczbę użytkowników, n - liczbę produktów na liście rekommendacji, u - użytkownika, k - pozycję produktu i na liście, g_{ui_k} jest poziomem zainteresowania (ang. *gain*) użytkownika u produktem i , zaś b jest parametrem przyjmującym wartość z zakresu [2,10] (zwykle $b = 2$).

nDCG (ang. *Normalized Cumulative Discounted Gain*)[2]

$$\text{nDCG} = \frac{\text{DCG}}{\text{DCG}^*} , \quad (15)$$

gdzie DCG^* jest idealnym DCG, w którym produkty na liście rekommendacyjnej są uporządkowane zgodnie z istotnością dla użytkownika.

Przykład 3 (c.d.)

$$\text{MRR} = \frac{1}{N} \sum_{l=1}^N \frac{1}{\text{rank}_l} = \frac{1}{1} \sum_{l=1}^1 \frac{1}{2} = \frac{1}{2}$$

$$\text{AP} = \frac{\sum_{k=1}^n (\text{prec}@k \cdot \text{ist}(k))}{|\{\text{istotne produkty}\}|} = \frac{0 * 0 + \frac{1}{2} * 1 + \frac{1}{3} * 0 + \frac{1}{4} * 0 + \frac{2}{5} * 1}{3} = \frac{3}{10}$$

$$\text{DCG} = \frac{1}{N} \sum_{u=1}^N \sum_{k=1}^n \frac{g_{ui_k}}{\max(1, \log_2 k)} = \left(\frac{0}{1} + \frac{4}{1} + \frac{0}{1,58} + \frac{0}{2} + \frac{5}{2,32} \right) \approx 6,16$$

$$\text{DCG}^* = \left(\frac{5}{1} + \frac{4}{1} + \frac{4}{1,58} + \frac{3}{2} + \frac{2}{2,32} \right) \approx 13,89$$

$$\text{nDCG} = \frac{\text{DCG}}{\text{DCG}^*} \approx \frac{6,16}{13,89} \approx 0,44$$

Przykład 4

Dokładny system rekomendacyjny

- Założymy, że mamy system rekomendujący produkty przed wejściem do supermarketu.
- Nasz system rekomenduje wszystkim zakupienie mleka i chleba.
- Ponad 90% osób wychodzących ze sklepu ma w siatce mleko i chleb.
- Czy system jest użyteczny?

Przykład 5

System rekomendujący muzykę

- Założymy, że mamy system rekomendujący kolejny utwór do playlisty.
- Nasz system rekomenduje zawsze jeden z pięciu utworów, które słuchamy najczęściej.
- Czy jesteśmy zadowoleni z tych rekomendacji?

- Nowość [1]

$$novelty = \frac{1}{k} \sum_{i \in R_{u,k}} \log_2(\text{pop}(i)) , \quad (16)$$

gdzie u oznacza użytkownika, k jest rozmiarem listy z rekommendacjami $R_{u,k}$, i oznacza produkt, zaś $\text{pop}(i)$ jest jego popularnością, t.j. liczbą użytkowników, którzy ocenili produkt i znormalizowaną przez liczbę użytkowników.

- Nowość [1]

$$\text{novelty} = \frac{1}{k} \sum_{i \in R_{u,k}} \log_2(\text{pop}(i)) , \quad (16)$$

gdzie u oznacza użytkownika, k jest rozmiarem listy z rekommendacjami $R_{u,k}$, i oznacza produkt, zaś $\text{pop}(i)$ jest jego popularnością, t.j. liczbą użytkowników, którzy ocenili produkt i znormalizowaną przez liczbę użytkowników.

- Różnorodność [3]

$$\text{ILD}(R) = \frac{1}{|R|(|R| - 1)} \sum_{i,j \in R, i \neq j} (1 - \text{sim}(i, j)) , \quad (17)$$

gdzie i, j są produktami, zaś sim dowolną miarą podobieństwa.

- Serendipity

- Serendipity
 - Przewidywalność (ang.) [6]

$$expectedness = \frac{1}{k} \sum_{i=1}^k pop(i) , \quad (18)$$

gdzie k jest rozmiarem listy z rekommendacjami $R_{u,k}$, i oznacza produkt, zaś $pop(i)$ jest jego popularnością.

- Serendipity
 - Przewidywalność (ang.) [6]

$$\text{expectedness} = \frac{1}{k} \sum_{i=1}^k \text{pop}(i) , \quad (18)$$

gdzie k jest rozmiarem listy z rekomendacjami $R_{u,k}$, i oznacza produkt, zaś $\text{pop}(i)$ jest jego popularnością.

- Unserendipity [5]

$$\text{unserendipity} = \frac{1}{|H_u|} \sum_{h \in H_u} \frac{1}{k} \sum_{i \in R_{u,k}} \text{sim}(i, h) , \quad (19)$$

gdzie u oznacza użytkownika, h jest produktem z historii użytkownika H_u (historyczne oceny użytkownika), k jest rozmiarem listy rekomendacji $R_{u,k}$ dla użytkownika u , i oznacza produkt z listy $R_{u,k}$, zaś sim dowolną miarą podobieństwa.

- Metody zaczerpnięte z uczenia maszynowego
 - Podział na zbiór uczący i testowy
 - w tym również podejście "pozostaw jeden" (ang. *leave-one-out*)
 - Walidacja skrośna

- Metody zaczerpnięte z uczenia maszynowego
 - Podział na zbiór uczący i testowy
 - w tym również podejście "pozostaw jeden" (ang. *leave-one-out*)
 - Walidacja skrośna
- Sposoby podziału zbioru:
 - po ocenach
 - po użytkownikach
 - po produktach

- Metody zaczerpnięte z uczenia maszynowego
 - Podział na zbiór uczący i testowy
 - w tym również podejście "pozostaw jeden" (ang. *leave-one-out*)
 - Walidacja skrośna
- Sposoby podziału zbioru:
 - po ocenach
 - po użytkownikach
 - po produktach
- Problem - nadmierne dopasowanie do danych
 - Rozwiązanie - dodatkowy zbiór walidacyjny

Co jest najważniejsze w ocenie jakości systemów rekomendacyjnych?

Najważniejsze przy ocenie jakości rekomendacji

- Znajomość dziedziny i problemu, który chcemy rozwiązać.
- Znajomość miar i świadomość, co tak naprawdę mierzą.
- Indywidualne podejście do każdego problemu.

Nie ma idealnej metody oceny jakości rekomendacji dla każdego przypadku, tak jak nie istnieje idealny algorytm rekomendacyjny!

 P. Castells and S. Vargas.

Novelty and diversity metrics for recommender systems: Choice, discovery and relevance.

In *In Proceedings of International Workshop on Diversity in Document Retrieval (DDR)*, pages 29–37, 2011.

 K. Järvelin and J. Kekäläinen.

Cumulated gain-based evaluation of ir techniques.

ACM Trans. Inf. Syst., 20(4):422–446, Oct. 2002.

 B. Smyth and P. McClave.

Similarity vs. diversity.

In D. W. Aha and I. Watson, editors, *Case-Based Reasoning Research and Development: 4th International Conference on Case-Based Reasoning, ICCBR 2001 Vancouver, BC, Canada, July 30 – August 2, 2001 Proceedings*, pages 347–361, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

 A. Turpin and F. Scholer.

User performance versus precision measures for simple search tasks.

In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pages 11–18, New York, NY, USA, 2006. ACM.

 Y. C. Zhang, D. O. Séaghdha, D. Quercia, and T. Jambor.

Auralist: Introducing serendipity into music recommendation.

In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, pages 13–22, New York, NY, USA, 2012. ACM.

 C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen.

Improving recommendation lists through topic diversification.

In Proceedings of the 14th International Conference on World Wide Web, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.

1. J. A. Konstan, M. D. Ekstrand. *Recommender Systems: Evaluation and Metrics*. University of Minnesota. Coursera.
2. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2010. *Recommender Systems Handbook* (1st ed.). Springer-Verlag New York, Inc., New York, NY, USA.
3. Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction* (1st ed.). Cambridge University Press, New York, NY, USA.
4. Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer International Publishing.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Ewaluacja systemów rekomendacyjnych z udziałem użytkowników

dr inż. Aleksandra Karpus

22 października 2021



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESA RADY MINISTRÓW

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)».

Podejścia do ewaluacji systemów rekomendacyjnych

- Retrospektywne - z wykorzystaniem danych historycznych
- Prospektywne - z udziałem użytkowników

Po co oceniać systemy rekomendacyjne z udziałem użytkowników?

- wyjaśni, w jaki sposób użytkownik korzysta z systemu.

- wyjaśni, w jaki sposób użytkownik korzysta z systemu.
- odpowie nam na pytanie, jak użytkownik odbiera wyświetlane mu informacje.

- wyjaśni, w jaki sposób użytkownik korzysta z systemu.
- odpowie nam na pytanie, jak użytkownik odbiera wyświetlane mu informacje.
- wyjaśni nam, jak użytkownik podejmuje decyzje.

Z jakich metod możemy skorzystać,
żeby ocenić system rekomendacyjny
z udziałem użytkowników?

- Logi aktywności
- Ankiety
- Zogniskowany wywiad grupowy
- Eksperymenty laboratoryjne/kontrolowane
- Testy A/B

Umożliwiają:

- analizę zachowania użytkownika podczas interakcji z systemem,

Umożliwiają:

- analizę zachowania użytkownika podczas interakcji z systemem,
- retrospektywną ocenę systemu rekomendacyjnego,

Umożliwiają:

- analizę zachowania użytkownika podczas interakcji z systemem,
- retrospektywną ocenę systemu rekomendacyjnego,
- wyznaczenie faktycznej dokładności algorytmu rekomendacyjnego.

Umożliwiają:

- analizę zachowania użytkownika podczas interakcji z systemem,
- retrospektywną ocenę systemu rekomendacyjnego,
- wyznaczenie faktycznej dokładności algorytmu rekomendacyjnego.

Są obiektywne!

Jakie informacje możemy logować?

- Dane użytkownika.
- Znacznik sesji.
- Co zostało wyświetcone użytkownikowi.
- W co użytkownik kliknął.
- Jak dużo czasu spędził na stronie produktu/wiadomości.
- Czy kupił produkt?
- Oznaczenia, jaki algorytm wygenerował daną rekomendację (dla metod hybrydowych).
- Itd.

Ograniczenia logów aktywności

- Ich użyteczność zależy od danych, które są logowane.

- Ich użyteczność zależy od danych, które są logowane.
- Nadal nie wiemy, jak użytkownik odbiera prezentowane mu informacje,

- Ich użyteczność zależy od danych, które są logowane.
- Nadal nie wiemy, jak użytkownik odbiera prezentowane mu informacje,
- ... ani w jaki sposób podejmuje decyzje.

Metody pozwalające na zapytanie użytkownika wprost, co myśli o systemie.

Metody pozwalające na zapytanie użytkownika wprost, co myśli o systemie.

Ankiety

- Przeprowadzana na losowej grupie osób.
- Polega na zadaniu użytkownikom serii pytań.
- Pytania muszą być odpowiednio przygotowane.
- Zwykle stosuje się odpowiedzi w formie pięcio- lub siedmio-stopniowej skali Likerta.

Metody pozwalające na zapytanie użytkownika wprost, co myśli o systemie.

Ankiety

- Przeprowadzana na losowej grupie osób.
- Polega na zadaniu użytkownikom serii pytań.
- Pytania muszą być odpowiednio przygotowane.
- Zwykle stosuje się odpowiedzi w formie pięcio- lub siedmio-stopniowej skali Likerta.

Zogniskowany wywiad grupowy

- Przeprowadzany na odpowiednio wybranej grupie.
- Polega na prezentowaniu fragmentów systemu i zbieraniu opinii.
- Nie można mylić zbierania informacji ze sprzedażą czegoś.

Wady ankiet/zogniskowanego wywiadu grupowego

- Użytkownicy nie muszą mówić nam wszystkiego.

Wady ankiet/zogniskowanego wywiadu grupowego

- Użytkownicy nie muszą mówić nam wszystkiego.
- Duży nakład pracy, aby stworzyć rzetelną ankietę, która uzyska wiarygodne odpowiedzi.

- Potrzebujemy przynajmniej prototypu systemu.
- Mogą się odbywać w laboratorium albo online.
- Kontrolujemy zmienną dla różnych użytkowników.
- Pozwalamy użytkownikowi na interakcję z systemem, później zadajemy pytania o doświadczenia.
- Logujemy wszystko.

- Wybieramy dwie zbliżone do siebie charakterystyką grupy użytkowników.
- Każda grupa korzysta z innej wersji systemu.
- Po okresie testów badamy interesujący nas wskaźnik w obu grupach.

Co jest najważniejsze w ocenie jakości systemów rekomendacyjnych?

Najważniejsze przy ocenie jakości rekomendacji

- Znajomość dziedziny i problemu, który chcemy rozwiązać.
- Indywidualne podejście do każdego problemu.

Nie ma idealnej metody oceny jakości rekomendacji dla każdego przypadku, tak jak nie istnieje idealny algorytm rekomendacyjny!

1. J. A. Konstan, M. D. Ekstrand. *Recommender Systems: Evaluation and Metrics*. University of Minnesota. Coursera.
2. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2010. *Recommender Systems Handbook* (1st ed.). Springer-Verlag New York, Inc., New York, NY, USA.
3. Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction* (1st ed.). Cambridge University Press, New York, NY, USA.
4. Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer International Publishing.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Filtrowanie oparte na zawartości

dr inż. Aleksandra Karpus

22 października 2021



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESA RADY MINISTRÓW

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

- Niespersonalizowane
- Spersonalizowane
 - **Filtrowanie oparte na zawartości (CBF)**
 - Filtrowanie kolaboracyjne (CF)
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Inne

- Ludzie mają swoje preferencje, np.:
 - amerykańskie filmy akcji
 - kuchnia azjatycka
 - książki napisane przez Stephena Kinga
 - czarne ubrania
 - itd.
- Ludzie częściej wybierają produkty podobne do tych, które już znają.

W jaki sposób możemy poznać preferencje użytkownika?

Sposoby pozyskiwania preferencji użytkownika

- Poprosić użytkownika, aby podał swoje preferencje.

Sposoby pozyskiwania preferencji użytkownika

- Poprosić użytkownika, aby podał swoje preferencje.
 - Niekoniecznie jest to dobra opcja.
 - Lepiej poprosić/umożliwić użytkownikowi edytować preferencje.

Sposoby pozyskiwania preferencji użytkownika

- Poprosić użytkownika, aby podał swoje preferencje.
 - Niekoniecznie jest to dobra opcja.
 - Lepiej poprosić/umożliwić użytkownikowi edytować preferencje.
- Wywnioskować na podstawie zachowania użytkownika (kliknięcia, zakup, itp.).
- Wywnioskować na podstawie ocen użytkownika.

Sposoby pozyskiwania preferencji użytkownika

- Poprosić użytkownika, aby podał swoje preferencje.
 - Niekoniecznie jest to dobra opcja.
 - Lepiej poprosić/umożliwić użytkownikowi edytować preferencje.
- Wywnioskować na podstawie zachowania użytkownika (kliknięcia, zakup, itp.).
- Wywnioskować na podstawie ocen użytkownika.
 - Dwa powyższe generują problem: w jaki sposób przekształcić kliknięcie bądź ocenę produktu w preferencje dotyczące cech produktów?

Sposoby pozyskiwania preferencji użytkownika

- Poprosić użytkownika, aby podał swoje preferencje.
 - Niekoniecznie jest to dobra opcja.
 - Lepiej poprosić/umożliwić użytkownikowi edytować preferencje.
- Wywnioskować na podstawie zachowania użytkownika (kliknięcia, zakup, itp.).
- Wywnioskować na podstawie ocen użytkownika.
 - Dwa powyższe generują problem: w jaki sposób przekształcić kliknięcie bądź ocenę produktu w preferencje dotyczące cech produktów?
- Można połączyć powyższe.

W jaki sposób możemy reprezentować preferencje użytkownika?

Naiwna reprezentacja preferencji użytkownika

- Możemy użyć słów kluczowych (cech) do opisu produktów.
- Do określenia preferencji użytkownika możemy użyć zwykłego zliczania słów kluczowych w produktach, które użytkownik lubił i których nie lubił.

- Użytkownik był zadowolony z obejrzenia 8 filmów akcji.
- Użytkownik odmówił obejrzenia 3 filmów z Melem Gibsonem.
- System rekomenduje użytkownikowi film "Zabójcza broń".
- Czy użytkownik będzie zadowolony z tej rekomendacji? Dlaczego?

- Metoda z dziedziny pozyskiwania informacji.
- $\text{TF} \times \text{IDF}$
- TF - częstotliwość słowa (ang. *term frequency*), np. zliczanie
- IDF (ang. *inverse document frequency*) - jak mało dokumentów zawiera dane słowo

$$\text{IDF} = \log\left(\frac{\#\text{dokumentów}}{\#\text{dokumentów zawierających słowo}}\right)$$

Co właściwie robi TF-IDF?

- Automatycznie degraduje istotność przyimków i innych pospolitych słów.
- Promuje istotne słowa bardziej od tych, które występują incydentalnie.

W jaki sposób możemy reprezentować produkty?

Przykład 2

reprezentacja filmów

| Tytuł | Reżyser | Gatunek | Aktor |
|----------------------|-------------------|-------------------------------------|--|
| Donnie Darko | Richard Kelly | drama, supernatural | Jake Gyllenhaal, Jena Malone, Drew Barrymore |
| Girl Interrupted | James Mangold | drama | Winona Ryder, Angelina Jolie |
| Inception | Christopher Nolan | heist, thriller, science fiction | Leonardo DiCaprio, Ken Watanabe, Marion Cotillard |
| Hunger Games | Gary Ross | science fiction, adventure | Jennifer Lawrence, Josh Hutcherson, Liam Hemsworth |
| Sleepless In Seattle | Nora Ephron | drama, comedy, romantic | Tom Hanks, Meg Ryan |

Przykład 2 (c.d.)

Reprezentacja filmów w formie macierzy

| Tytuł | Richard Kelly | James Mangold | ... | drama | science fiction | ... | Meg Ryan |
|----------------------|---------------|---------------|-----|-------|-----------------|-----|----------|
| Donnie Darko | 1 | 0 | ... | 1 | 0 | ... | 0 |
| Girl Interrupted | 0 | 1 | ... | 1 | 0 | ... | 0 |
| Inception | 0 | 0 | ... | 0 | 1 | ... | 0 |
| Hunger Games | 0 | 0 | ... | 0 | 1 | ... | 0 |
| Sleepless In Seattle | 0 | 0 | ... | 1 | 0 | ... | 1 |

Macierz zawartości produktu

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{11} & c_{12} & \cdots & c_{1m} \\ \cdots & \cdots & \ddots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix}$$

- Wiersze reprezentują produkty,
- Kolumny reprezentują cechy/atrzybuty produktu,
- n – liczba produktów, m – liczba cech (atrzybutów) produktu,
- c_{ia} – indykator, czy produkt i zawiera atrybut a .

Jakie wartości może przyjmować c_{ia} ?

- wartości ze zbioru {0, 1}
- ilość wystąpień danej cechy/słowa kluczowego w produkcie/dokumencie
- TF-IDF

Przykład 3

Wyznaczanie wartości TF-IDF

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| i | 1 | 0 | 1 | 0 | 0 | 1 |
| j | 1 | 1 | 1 | 1 | 1 | 1 |
| x | 1 | 0 | 0 | 1 | 0 | 1 |
| y | 1 | 0 | 1 | 0 | 0 | 0 |

$$\text{TF}_{a,i} = \frac{N_{a,i}}{N_i}$$

- $N_{a,i}$ - liczba wystąpień atrybutu a w produkcie i
- N_i - liczba atrybutów produktu i

Przykład 3 (c.d.)

Wyznaczanie wartości TF

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| i | 1 | 0 | 1 | 0 | 0 | 1 |
| j | 1 | 1 | 1 | 1 | 1 | 1 |
| x | 1 | 0 | 0 | 1 | 0 | 1 |
| y | 1 | 0 | 1 | 0 | 0 | 0 |

$$TF_{a,i} = \frac{N_{a,i}}{N_i} = \frac{1}{3}$$

$$TF_{b,i} = \frac{N_{b,i}}{N_i} = \frac{0}{3} = 0$$

$$TF_{a,j} = \frac{N_{a,j}}{N_j} = \frac{1}{6}$$

Przykład 3 (c.d.)

Wyznaczanie wartości IDF

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| i | 1 | 0 | 1 | 0 | 0 | 1 |
| j | 1 | 1 | 1 | 1 | 1 | 1 |
| x | 1 | 0 | 0 | 1 | 0 | 1 |
| y | 1 | 0 | 1 | 0 | 0 | 0 |

$$\text{IDF}_a = \log \frac{N}{N_a}$$

- N - liczba wszystkich produktów
- N_a - liczba produktów posiadających atrybut a

Przykład 3 (c.d.)

Wyznaczanie wartości IDF (2)

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| i | 1 | 0 | 1 | 0 | 0 | 1 |
| j | 1 | 1 | 1 | 1 | 1 | 1 |
| x | 1 | 0 | 0 | 1 | 0 | 1 |
| y | 1 | 0 | 1 | 0 | 0 | 0 |

$$\text{IDF}_a = \log \frac{N}{N_a} = \log \frac{4}{4} = 0,00$$

$$\text{IDF}_b = \log \frac{N}{N_b} = \log \frac{4}{1} = 0,60$$

$$\text{IDF}_c = \log \frac{N}{N_c} = \log \frac{4}{3} = 0,12$$

Przykład 4

Ulepszenie macierzy zawartości produktu

| Tytuł | Reżyser | Gatunek | Aktor |
|----------------------|-------------------|-------------------------------------|--|
| Donnie Darko | Richard Kelly | drama, supernatural | Jake Gyllenhaal, Jena Malone, Drew Barrymore |
| Girl Interrupted | James Mangold | drama | Winona Ryder, Angelina Jolie |
| Inception | Christopher Nolan | heist, thriller, science fiction | Leonardo DiCaprio, Ken Watanabe, Marion Cotillard |
| Hunger Games | Gary Ross | science fiction, adventure | Jennifer Lawrence, Josh Hutcherson, Liam Hemsworth |
| Sleepless In Seattle | Nora Ephron | drama, comedy, romantic | Tom Hanks, Meg Ryan |

Przykład 4 (c.d.)

Ulepszenie macierzy zawartości produktu

| Tytuł | Reżyser = 0,75 | Gatunek = 1,00 | Aktor = 0,50 |
|----------------------|-------------------|-------------------------------------|--|
| Donnie Darko | Richard Kelly | drama, supernatural | Jake Gyllenhaal, Jena Malone, Drew Barrymore |
| Girl Interrupted | James Mangold | drama | Winona Ryder, Angelina Jolie |
| Inception | Christopher Nolan | heist, thriller, science fiction | Leonardo DiCaprio, Ken Watanabe, Marion Cotillard |
| Hunger Games | Gary Ross | science fiction, adventure | Jennifer Lawrence, Josh Hutcherson, Liam Hemsworth |
| Sleepless In Seattle | Nora Ephron | drama, comedy, romantic | Tom Hanks, Meg Ryan |

Przykład 4 (c.d.)

Ulepszenie macierzy zawartości produktu

| Tytuł | Reżyser = 0,75 | Gatunek = 1,00 | Aktor = 0,50 |
|----------------------|-------------------|-------------------------------------|--|
| Donnie Darko | Richard Kelly | drama, supernatural | Jake Gyllenhaal, Jena Malone, Drew Barrymore |
| Girl Interrupted | James Mangold | drama | Winona Ryder, Angelina Jolie |
| Inception | Christopher Nolan | heist, thriller, science fiction | Leonardo DiCaprio , =1.00 Ken Watanabe, Marion Cotillard |
| Hunger Games | Gary Ross | science fiction, adventure | Jennifer Lawrence, Josh Hutcherson, Liam Hemsworth |
| Sleepless In Seattle | Nora Ephron | drama, comedy, romantic | Tom Hanks, Meg Ryan |

filtrowania opartego na zawartości

- Każdy atrybut jest wymiarem w pewnej przestrzeni.
- Każdy produkt można umieścić w tej przestrzeni. Miejsce to jest opisane przez pewien wektor.
- Preferencje użytkownika również można zapisać w postaci wektora w tej przestrzeni.
- Dopasowanie pomiędzy użytkownikiem i produktem jest wyznaczane na podstawie odległości tych wektorów w przestrzeni.

- Podobieństwo cosinusowe

$$\cos(u, v) = \frac{u \bullet v}{\|u\| \|v\|} , \quad (1)$$

gdzie \bullet oznacza iloczyn skalarny, zaś $\|u\|$ jest normą wektora u .

- W przypadku wektorów binarnych bez wag można skorzystać z podobieństwa Jaccarda dla zbiorów:

$$jacc(A, B) = \frac{|A \cap B|}{|A \cup B|} , \quad (2)$$

gdzie \cap oznacza iloczyn, a \cup sumę zbiorów A i B .

- W jaki sposób tworzyć profil użytkownika?
 - Agregować produkty z historii użytkownika?
 - W jaki sposób to robić?

Sposoby agregacji produktów przy tworzeniu profilu użytkownika

- Dla danych *implicite* - agregacja produktów bez wag,
- Dla danych typu "lubi/nie lubi" - dodać lubiane produkty i odjąć te nielubiane,
- Dla ocen:
 - Można ustawić pewien próg i agregować tylko produkty z oceną powyżej progu (bez uwzględniania wag),
 - Można użyć ocen jako wag - tylko dodatnie wagi,
 - Można wykorzystać dodatnie i ujemne wagi - znormalizowana skala ocen.

W jaki sposób robimy predykcję/ranking?

Wyzwania i wady metod opartych na zawartości

- Bazują na dobrze opisanych atrybutach, które korespondują z preferencjami użytkownika.
- Wymagają sensownego rozkładu atrybutów względem produktów.
- Raczej znajdują substytuty niż dopełnienia.
- Problemy:
 - Nadmierne dopasowanie do danych.
 - Problem nowego użytkownika.

Zalety metod opartych na zawartości

- Nie wymagają dużej liczby użytkowników.
- Świecznie się sprawdzają przy poszukiwaniu zamienników produktów, np. podobna alternatywa dla drogiego aparatu.
- Łatwo wyjaśnić użytkownikowi, skąd taka rekomendacja.

1. J. A. Konstan, M. D. Ekstrand. *Introduction to Recommender Systems*. University of Minnesota. Coursera.
2. P. Cremonesi. *Basic Recommender Systems*. EIT Digital & Politecnico di Milano. Coursera.
3. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2010. *Recommender Systems Handbook* (1st ed.). Springer-Verlag New York, Inc., New York, NY, USA.
4. Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction* (1st ed.). Cambridge University Press, New York, NY, USA.
5. Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer International Publishing.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Zaawansowane metody filtrowania kolaboracyjnego

dr inż. Aleksandra Karpus

29 października 2021



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESA RADY MINISTRÓW
MINISTERSTWO ROZWOJU REGIONALNEGO

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

- Niespersonalizowane
- Spersonalizowane
 - Filtrowanie oparte na zawartości (CBF)
 - **Filtrowanie kolaboracyjne (CF)**
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Inne

Inny podział metod filtrowania kolaboracyjnego

- Bazujące na pamięci – wykonują obliczenia z wykorzystaniem całej macierzy ocen użytkowników,
- Bazujące na modelu – budują model, najczęściej na podstawie macierzy ocen użytkowników, i wykorzystują go do wykonywania obliczeń.

$$model = f(R)$$

$$\hat{r}_{ui} = g(model, profil_u)$$

Czym różnią się te dwa podejścia?

Różnice pomiędzy metodami bazującymi na pamięci i na modelu

- Metody bazujące na pamięci nie są w stanie wyznaczyć rekomendacji dla nowego użytkownika bez przeliczenia modelu.
- Metody bazujące na modelu nie muszą go przeliczać w przypadku pojawienia się nowego użytkownika, jeśli macierz R jest wystarczająco duża.

Algorytm **UserKNN** jest metodą
bazującą na pamięci czy na modelu?
Dlaczego?

Algorytm **ItemKNN** jest metodą
bazującą na pamięci czy na modelu?
Dlaczego?

- Niespersonalizowane
- Spersonalizowane
 - Filtrowanie oparte na zawartości (CBF)
 - **Filtrowanie kolaboracyjne (CF)**
 - User-based
 - **Item-based**
 - Rozkład macierzy
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Inne

Filtrowanie kolaboracyjne oparte na produkcie – przypomnienie

$$s(u, i) = \frac{\sum_{j \in J} r_{uj} * w_{ij}}{\sum_{j \in J} w_{ij}},$$

gdzie r_{uj} to ocena, którą użytkownik u dał produktowi j , w_{ij} to podobieństwo produktów i i j , zaś $s(u, i)$ to przewidywana ocena, jaką użytkownik u da produktowi i .

Filtrowanie kolaboracyjne – przekształcenia

$$s(u, i) = \frac{\sum_{j \in J} r_{uj} * w_{ij}}{\sum_{j \in J} w_{ij}}$$
$$\iff$$

$$\hat{r}_{ui} = \sum_{j \in J} r_{uj} * p_{ij}$$

Filtrowanie kolaboracyjne – przekształcenia (c.d.)

$$s(u, i) = \frac{\sum_{j \in J} r_{uj} * w_{ij}}{\sum_{j \in J} w_{ij}}$$
$$\iff$$

$$\hat{r}_{ui} = \sum_{j \in J} r_{uj} * p_{ij}$$
$$\iff$$

$$\hat{R} = R \cdot P$$

Systemy rekomendacyjne jako problem optymalizacyjny

- Wyznaczanie podobieństwa na podstawie macierzy ocen użytkowników nie jest zbyt efektywne.
- Arbitralnie wybrana miara podobieństwa niekoniecznie musi być najlepszą możliwą opcją.

- Miary błędu
 - MAE
 - MSE
 - RMSE
- Miary z dziedziny pozyskiwania informacji
 - Precyza
 - Czułość
 - AUC ROC
- Miary specyficzne dla rankingu
 - MRR
 - MAP
 - nDCG

Definicja problemu optymalizacji dla funkcji kosztu MSE

Problem rekomendacji:

$$\hat{R} = R \cdot P$$

Funkcja kosztu – MSE:

$$E(P) = \|R - RP\|_2^2$$

Szukamy P^* takiego, że:

$$E(P^*) = \min_P \|R - RP\|_2^2$$

"Najlepsze" rozwiązanie

$$P^* = I$$

ALE...

"Najlepsze" rozwiązanie

$$P^* = I$$

ALE...

brak generalizacji modelu!

- Wymuszenie $\text{diag}(P) = 0$
- Regularyzacja

$$E(P) = \|R - RP\|_2^2 + \lambda \|P\|_2 \quad , \quad \lambda > 0$$

Dla skrajnych wartości λ :

- $\lim_{\lambda \rightarrow 0} E(P) = \|R - RP\|_2^2 \Rightarrow P = I$
- $\lim_{\lambda \rightarrow \infty} E(P) = \lambda \|P\|_2 \Rightarrow P \rightarrow 0$

Sparse Linear Method

$$\hat{r}_{ui} = \mathbf{r}_u^T \mathbf{w}_i \quad , \quad (1)$$

gdzie $r_{ui} = 0$.

$$\begin{aligned} \min_{\mathbf{w}_j} \quad & \frac{1}{2} \|\mathbf{r}_j - \mathbf{R}\mathbf{w}_j\|_2^2 + \frac{\beta}{2} \|\mathbf{w}_j\|_2^2 + \lambda \|\mathbf{w}_j\|_1 \\ \mathbf{w}_j \geqslant & 0 \\ w_{jj} = & 0 \end{aligned} \quad (2)$$

- łatwy do zrównoleglenia obliczeń,
- korzysta z regularizacji "elastyczna sieć".

- Niespersonalizowane
- Spersonalizowane
 - Filtrowanie oparte na zawartości (CBF)
 - **Filtrowanie kolaboracyjne (CF)**
 - User-based
 - Item-based
 - **Rozkład macierzy**
 - Kontekstowe (CARS)
 - Wielodziedzinowe (CDRS)
 - Inne

Algebra liniowa – macierze

Przypomnienie

Wyznacznik macierzy

- Ma sens tylko dla macierzy kwadratowych.
- Jest dany wzorem rekurencyjnym:

$$\det [a] = a$$

$$\det \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \ddots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = a_{11} \det \begin{bmatrix} a_{22} & \cdots & a_{2n} \\ \cdots & \ddots & \cdots \\ a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$-a_{12} \det \begin{bmatrix} a_{21} & \cdots & a_{2n} \\ \cdots & \ddots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} + \cdots \pm a_{1n} \det \begin{bmatrix} a_{21} & \cdots & a_{2(n-1)} \\ \cdots & \ddots & \cdots \\ a_{n1} & \cdots & a_{n(n-1)} \end{bmatrix}$$

Przykład 1

Wyznaczanie wyznacznika macierzy

$$\begin{bmatrix} 1 & -1 & 2 \\ 0 & 2 & 1 \\ 2 & -1 & 0 \end{bmatrix}$$

Definicja

Skalar $\lambda \in \mathbb{R}$ nazywamy **wartością własną** macierzy $A \in \mathbb{R}^{n \times n}$ jeżeli istnieje wektor $v \neq 0$, taki że

$$Av = \lambda v.$$

Wektor v nazywamy **wektorem własnym** odpowiadającym wartości własnej λ .

Definicja

Skalar $\lambda \in \mathbb{R}$ nazywamy **wartością własną** macierzy $A \in \mathbb{R}^{n \times n}$ jeżeli istnieje wektor $v \neq 0$, taki że

$$Av = \lambda v.$$

Wektor v nazywamy **wektorem własnym** odpowiadającym wartości własnej λ .

Twierdzenie

Dla macierzy $A \in \mathbb{R}^{n \times n}$ następujące warunki są równoważne:

1. λ jest wartością własną A ;
2. układ równań $(A - \lambda I)v = 0$ ma niezerowe rozwiązanie;
3. $\det(A - \lambda I) = 0$.

Wartości osobliwe σ_i macierzy A to pierwiastki z wartości własnych macierzy, tj. $\lambda_i = \sigma_i^2$.

Rozkład macierzy według wartości osobliwych

Dla każdej macierzy rzeczywistej A istnieje rozkład:

$$A = U\Sigma V^T,$$

gdzie:

- Macierze U i V^T są ortogonalne,
- Macierz Σ jest kwadratową macierzą diagonalną, której wartości σ_i są nieujemnymi wartościami osobliwymi macierzy A zwyczajowo uporządkowanymi nierośnąco.
- Macierz A ma wymiar $n \times m$, $U - n \times k$, $V - m \times k$, a $\Sigma - k \times k$.

Prezentacja graficzna

$$A = U \Sigma V^T$$

Diagram illustrating the Singular Value Decomposition (SVD) of a matrix A. The matrix A is shown in a green box with dimensions mxn below it. To its right is an equals sign. Following the equals sign are three boxes: U in light blue with dimensions mxk, Sigma in yellow with dimensions kxk, and V^T in orange with dimensions kxn.

Wyznaczanie rozkładu macierzy A (na tablicy).

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

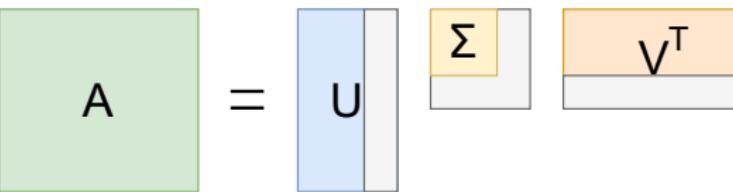
Motywacje zastosowania rozkładu macierzy w systemach rekomendacyjnych

- Macierz ocen jest zbyt specyfczną reprezentacją preferencji użytkownika, która może prowadzić do problemów z generalizacją.
- Ludzie zwykle określają swoje zainteresowania i preferencje poprzez cechy i opis produktów, a nie poprzez przykładowe produkty.

Interpretacja

$$A = U \Sigma V^T$$

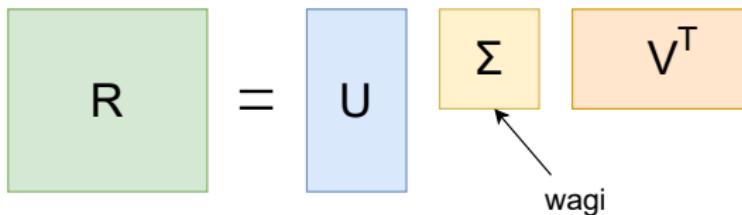
mxn mxd dxd dxn



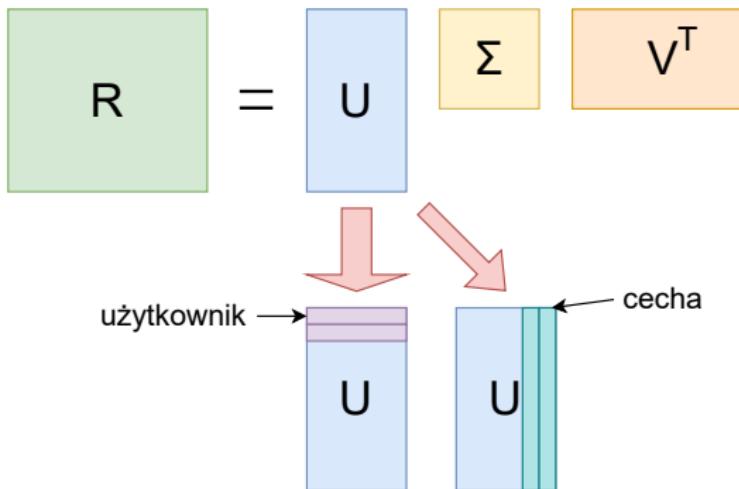
Interpretacja (c.d.)

$$R = U \Sigma V^T$$

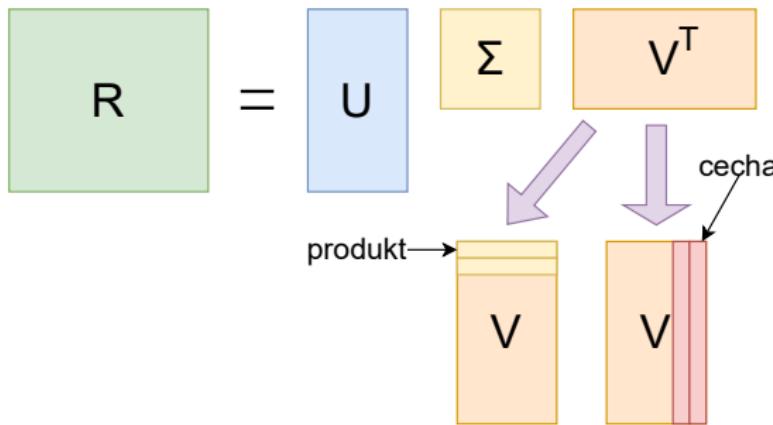
wagi



Interpretacja (c.d.)



Interpretacja (c.d.)



Rozkład macierzy a systemy rekommendacyjne

- SVD reprezentuje preferencje w formie cech ukrytych.
- Cechy te są uczone na podstawie ocen użytkowników → raczej nie są interpretowalne przez człowieka.
- Ponieważ wagi σ_i są uporządkowane nierosnąco, możemy ograniczyć wymiar k i mieć najlepsze możliwe przybliżenie macierzy R o zredukowanym wymiarze $d \ll \min\{m, n\}$.

Algorytm SVD – formalizacja

Singular Value Decomposition

- Model łączy każdego użytkownika u z wektorem preferencji użytkownika $\mathbf{p}_u \in \mathbb{R}^f$.
- Macierz P jest tutaj równa $U\Sigma$.
- Model łączy każdy produkt i z wektorem cech ukrytych $\mathbf{q}_i \in \mathbb{R}^f$.
- Iloczyn skalarny wektorów $\mathbf{q}_i^T \mathbf{p}_u$ wychwytuje całkowite zainteresowanie użytkownika u w charakterystyce przedmiotu i .
- Przewidywana ocena użytkownika u dla przedmiotu i jest wyliczana ze wzoru:

$$\hat{r}_{ui} = \mu + b_i + b_u + \mathbf{q}_i^T \mathbf{p}_u , \quad (3)$$

gdzie μ jest średnią globalną ze wszystkich ocen w macierzy ocen, zaś b_i i b_u są odpowiednio stronniczością produktu i użytkownika.

- SVD++ jest rozszerzeniem algorytmu SVD.
- Poza znajdywaniem zależności pomiędzy użytkownikiem i produktami w przestrzeni cech ukrytych, rozpatruje również wszystkie interakcje użytkownika z produktami (ang. *implicit feedback*).
- Przewidywana ocena użytkownika u dla przedmiotu i jest wyliczana ze wzoru:

$$\hat{r}_{ui} = \mu + b_i + b_u + \mathbf{q}_i^T \left(\mathbf{p}_u + |R(u)|^{\frac{1}{2}} \sum_{j \in R(u)} y_j \right) , \quad (4)$$

gdzie symbole oznaczają to samo, co dla równania (3), $R(u)$ to produkty z historii użytkownika u , zaś wektor y_j jest reprezentacją produktu j z historii $R(u)$ w przestrzeni cech ukrytych.

Weighted Regularized Matrix Factorization

- Metoda bazuje na binarnej skali ocen: *lubi* i *nie lubi*.
- Istniejąca skala ocen jest zamieniana na 1 – "lubi" i 0 – "nie lubi" zgodnie ze wzorem:

$$p_{ui} = \begin{cases} 1, & r_{ui} > 0 \\ 0, & r_{ui} = 0 \end{cases} . \quad (5)$$

- Ponieważ preferencje są wyuczane, wprowadza się poziomy ufności poprzez zmienne:

$$c_{ui} = 1 + \alpha r_{ui} , \quad (6)$$

gdzie α jest stałą zależną od zbioru danych.

- Również w tej metodzie preferencje użytkownika dotyczące produktów są reprezentowane poprzez iloczyn skalarny wektorów cech ukrytych:

$$p_{ui} = \mathbf{x}_u^T \mathbf{y}_i . \quad (7)$$

- W celu znalezienia wektorów cech ukrytych rozwiązuje się poniższy problem optymalizacyjny:

$$\min_{x^*, y^*} \quad \sum_{u,i} c_{ui} (p_{ui} - \mathbf{x}_u^T \mathbf{y}_i)^2 + \lambda \left(\sum_u \|\mathbf{x}_u\|^2 + \sum_i \|\mathbf{y}_i\|^2 \right) , \quad (8)$$

gdzie λ jest parametrem regularyzacji.

- Dobór ilości wymiarów d .
- Rekomendacje są trudne do uzasadnienia użytkownikowi – bazują na wyliczonych cechach ukrytych, które są niezrozumiałe dla człowieka.
- Problemy metod bazujących na pamięci.

Algorytm BPR jako przykład metody uczenia rankingu

- Dobry system rekommendacyjny nie musi umieć przewidywać oceny użytkownika dla produktu.
- Ważne jest, aby system potrafił przewidzieć preferencje użytkownika dotyczące nieznanego produktu i zwracał listę kilku satysfakcjonujących użytkownika produktów.

Relację $>_u$ nazywamy **całkowitym porządkiem**, jeżeli spełnia następujące własności:

$$\forall i, j \in I : i \neq j \Rightarrow i >_u j \vee j >_u i , \quad (9)$$

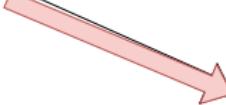
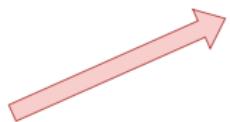
$$\forall i, j \in I : i >_u j \wedge j >_u i \Rightarrow i = j , \quad (10)$$

$$\forall i, j, k \in I : i >_u j \wedge j >_u k \Rightarrow i >_u k . \quad (11)$$

- Algorytm został zaproponowany dla danych "nie wprost".
- W tej metodzie interesuje nas, który z dwóch produktów użytkownik woli.
- Zakłada się, że użytkownik preferuje produkt, który zna, od tego, którego nie zna.
- Na podstawie macierzy interakcji nie jesteśmy w stanie powiedzieć, który z dwóch znanych (bądź dwóch nieznanych) produktów użytkownik woli.
- Relację preferencji oznaczamy symbolem $>_u$ i wymagamy, aby spełniała warunki całkowitego porządku.
- Zapis $i >_u j$ oznacza, że użytkownik u woli produkt i od produktu j .

Graficzna reprezentacja preferencji użytkownika

| | i_1 | i_2 | i_3 | i_4 |
|-------|-------|-------|-------|-------|
| u_1 | 0 | 1 | 1 | 0 |
| u_2 | 1 | 0 | 0 | 1 |
| u_3 | 1 | 1 | 0 | 0 |
| u_4 | 0 | 0 | 1 | 1 |
| u_5 | 0 | 0 | 1 | 0 |



$$u_1: i >_{u_1} j$$

$$i_1 \quad i_2 \quad i_3 \quad i_4$$

| j_1 | | + | + | ? |
|-------|---|---|---|---|
| j_2 | - | | ? | - |
| j_3 | - | ? | | - |
| j_4 | ? | + | + | |

...

$$u_5: i >_{u_5} j$$

$$i_1 \quad i_2 \quad i_3 \quad i_4$$

| j_1 | ? | + | ? |
|-------|---|---|---|
| j_2 | ? | | ? |
| j_3 | - | - | |
| j_4 | ? | ? | + |

Rysunek: Zamiana macierzy interakcji na relację preferencji $>_u$ (na podstawie [5]).

Bayesian Personalized Ranking

Algorytm BPR sprowadza się do rozwiązania problemu optymalizacyjnego

$$\max P(\hat{r}_{ui} > \hat{r}_{uj} | u)$$

przy założeniu, że:

$$P(\hat{r}_{ui} > \hat{r}_{uj} | u) = \sigma(x_{u_{ij}}),$$

gdzie:

$$\sigma(x_{u_{ij}}) = \frac{1}{1 + e^{-x_{u_{ij}}}},$$

$$x_{u_{ij}} = \hat{r}_{ui} - \hat{r}_{uj}$$

Algorytm BPR – podsumowanie

- Skupiamy się na optymalizacji uszeregowania par.
- Rozwiążując problem optymalizacyjny BPR, maksymalizujemy jednocześnie miarę AUC ROC.
- Jest to przykład algorytmu uczenia ranking (ang. *learning to rank*).

Bibliografia

1. A. Romanowski. *Algebra liniowa*. Wydawnictwo Politechniki Gdańskiej, 2003.
2. X. Ning and G. Karypis. *Slim: Sparse linear methods for top-n recommender systems*. In 2011 IEEE 11th International Conference on Data Mining, pages 497–506, Dec 2011.
3. Y. Koren and R. Bell. *Advances in Collaborative Filtering*, pages 145–186. Springer US, Boston, MA, 2011.
4. Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, pages 426–434, New York, NY, USA, 2008. ACM.
5. S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, pages 452–461, Arlington, Virginia, United States, 2009. AUAI Press.
6. Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08, pages 263–272, Washington, DC, USA, 2008. IEEE Computer Society.
7. Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer International Publishing.
8. J. A. Konstan, M. D. Ekstrand. *Introduction to Recommender Systems*. University of Minnesota. Coursera.
9. P. Cremonesi. *Advanced Recommender Systems*. EIT Digital & Politecnico di Milano. Coursera.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Kontekstowe systemy rekomendacyjne

dr inż. Aleksandra Karpus

15 listopada 2023



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESA RADY MINISTRÓW
MINISTERSTWA ROZWOJU ECONOMIC

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



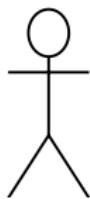
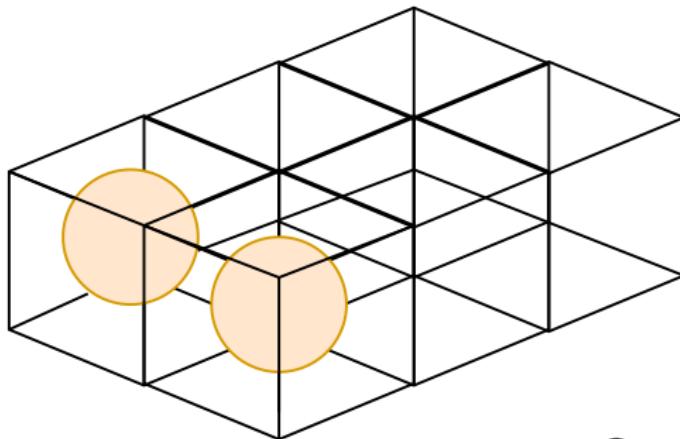
Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)».

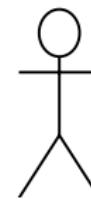
- Niespersonalizowane
- Spersonalizowane
 - Filtrowanie oparte na zawartości (CBF)
 - Filtrowanie kolaboracyjne (CF)
 - User-based
 - Item-based
 - Rozkład macierzy
 - Kontekstowe (**CARS**)
 - Wielodziedzinowe (CDRS)
 - Inne

Czym jest kontekst?

Przykład - magiczne pudełko¹

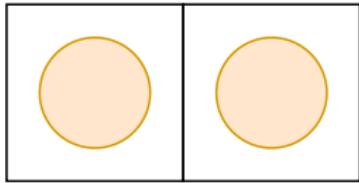


Adam

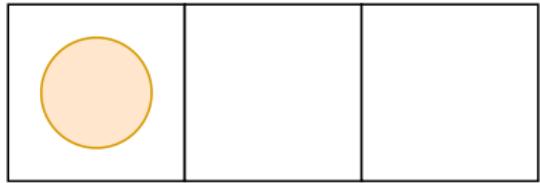


Basia

¹Na podstawie [10]

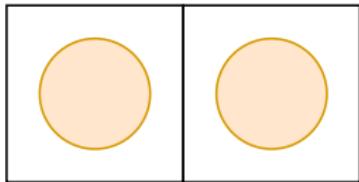


Widok Adama

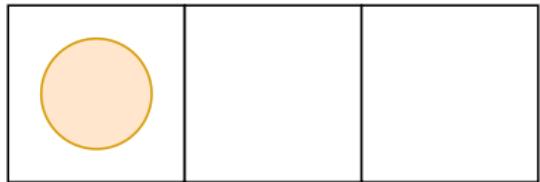


Widok Basi

²Na podstawie [10]



Widok Adama



Widok Basi

Kto ma rację?

²Na podstawie [10]

$$P_1 = V_1, \dots, P_n = V_n$$

Zdanie₁

Zdanie₂

...

gdzie P_1, \dots, P_n to parametry kontekstowe, zaś V_1, \dots, V_n to wartości tych parametrów.

Przykładowe zdania z parametrami

- Jestem tutaj na wykładzie z Odkrywania Wiedzy i Systemów Rekomendacyjnych.

- Jestem tutaj na wykładzie z Odkrywania Wiedzy i Systemów Rekomendacyjnych.
 - Parametry: podmiot i miejsce.

- Jestem tutaj na wykładzie z Odkrywania Wiedzy i Systemów Rekomendacyjnych.
 - Parametry: podmiot i miejsce.
 - Wartości: podmiot = "Aleksandra Karpus", miejsce = "sala 31 EA".

- Jestem tutaj na wykładzie z Odkrywania Wiedzy i Systemów Rekomendacyjnych.
 - Parametry: podmiot i miejsce.
 - Wartości: podmiot = "Aleksandra Karpus", miejsce = "sala 31 EA".
- Następny wykład odbędzie się za tydzień.

- Jestem tutaj na wykładzie z Odkrywania Wiedzy i Systemów Rekomendacyjnych.
 - Parametry: podmiot i miejsce.
 - Wartości: podmiot = "Aleksandra Karpus", miejsce = "sala 31 EA".
- Następny wykład odbędzie się za tydzień.
 - Parametry: czas.

- Jestem tutaj na wykładzie z Odkrywania Wiedzy i Systemów Rekomendacyjnych.
 - Parametry: podmiot i miejsce.
 - Wartości: podmiot = "Aleksandra Karpus", miejsce = "sala 31 EA".
- Następny wykład odbędzie się za tydzień.
 - Parametry: czas.
 - Wartości: czas = "15.11.2023".

Czy parametry kontekstowe to tylko podmiot, miejsce i czas?

- osoba
- miejsce
- czas
- pogoda
- towarzystwo
- wyznanie
- status majątkowy
- stan cywilny
- nastrój
- ...
- i wiele więcej

Czy te same parametry kontekstowe są istotne w systemach informatycznych/rekomendacyjnych?

Czy te same parametry kontekstowe są istotne w systemach informatycznych/rekomendacyjnych?

Dlaczego?

Przykłady parametrów kontekstowych typowych dla systemów informatycznych

- urządzenie
- system operacyjny
- rodzaj aplikacji
- sposób przechowywania danych
- ...

Czym jest kontekst w systemach informatycznych/rekomendacyjnych?

Kontekst to dowolna informacja, która może być wykorzystana do scharakteryzowania sytuacji encji. Encją może być osoba, miejsce lub obiekt/przedmiot, które są uznane za istotne z punktu widzenia interakcji pomiędzy użytkownikiem i aplikacją, włączając w to także samego użytkownika i aplikację.[3]

Co to znaczy, że system jest **kontekstowy**?

System nazywamy kontekstowym, jeżeli wykorzystuje on kontekst do dostarczenia użytkownikowi istotnych informacji i/lub usług, gdzie istotność zależy od zadań użytkownika.[4]

Czym więc są kontekstowe systemy rekomendacyjne?

Kontekstowe systemy rekomendacyjne

(ang. *Context-Aware Recommender Systems*, CARS) to rodzaj systemów rekomendacyjnych, który wykorzystuje informację o kontekście użytkownika, aby wygenerować lepsze rekomendacje.

Kontekstowe systemy rekomendacyjne

(ang. *Context-Aware Recommender Systems*, CARS) to rodzaj systemów rekomendacyjnych, który wykorzystuje informację o kontekście użytkownika, aby wygenerować lepsze rekomendacje.

$$R : \text{Użytkownicy} \times \text{Produkty} \times \text{Kontekst} \mapsto \text{Oceny} , \quad (1)$$

gdzie *Kontekst* oznacza informację kontekstową istotną z punktu widzenia dziedziny zastosowań.

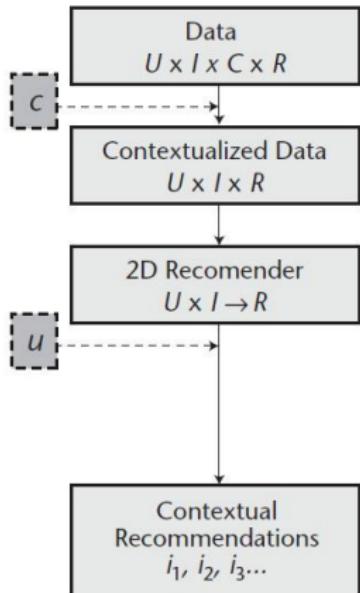
- Ludzie w różnych sytuacjach wybierają różne rzeczy, np. inne wybieramy książki dla rozrywki, a inne do pracy.
- Preferencje mogą się zmieniać z czasem.
- Możemy kupować rzeczy dla kogoś innego.

Rodzaje kontekstowych systemów rekommendacyjnych

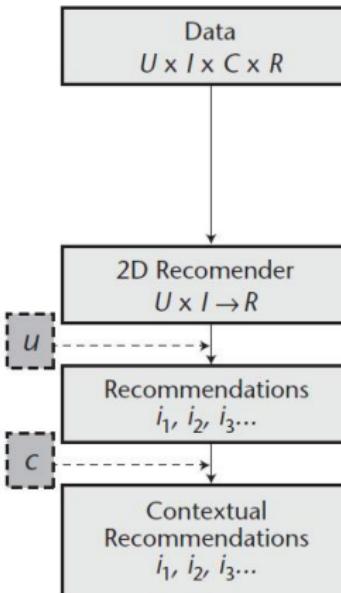
- kontekstowa pre-filtracja (ang. *contextual pre-filtering*)
- kontekstowa post-filtracja (ang. *contextual post-filtering*)
- modelowanie kontekstowe (ang. *contextual modeling*)

Rodzaje kontekstowych systemów rekomendacyjnych (c.d.)

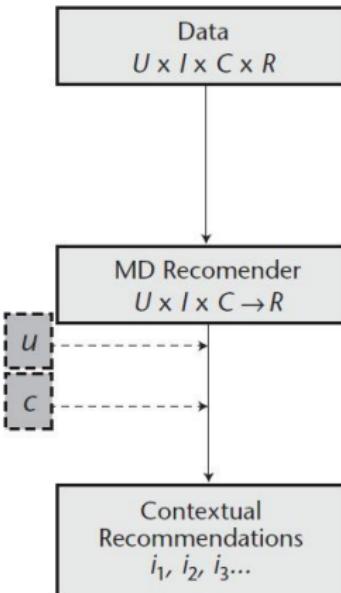
(a) Contextual Prefiltering



(b) Contextual Postfiltering



(c) Contextual Modeling



Rysunek: Zaczepnięte z [1].

Czy zawsze powinniśmy używać wszystkich dostępnych parametrów kontekstowych?

Czy zawsze powinniśmy używać wszystkich dostępnych parametrów kontekstowych?

Dlaczego?

W jakich sytuacjach parametr nie jest dobrym parametrem kontekstowym?

Cechy źle dobranych parametrów kontekstowych

- Nieodpowiedni dla dziedziny zastosowań.
- Ma stałą wartość dla większości przypadków.
- Brak związku pomiędzy wartością parametru a badaną cechą, np. oceną użytkownika.

W jaki sposób wybrać dobre parametry kontekstowe?

Badanie zmienności parametru kontekstowego³ (1)

- entropia

$$\eta_e(v) = -\frac{1}{\log_2(M)} \sum_{i=1}^M \frac{n_i}{n} \log_2 \frac{n_i}{n}, \quad (2)$$

gdzie M jest liczbą możliwych wartości zmiennej kontekstowej v , n_i jest liczbą wystąpień wartości i dla zmiennej v , zaś n jest liczbą wszystkich obserwacji zmiennej v . $1/\log 2M$ jest stałą normalizującą, która zapewnia nam, że wartości entropii zawsze będą pomiędzy 0 a 1.

³[8]

Badanie zmienności parametru kontekstowego ⁴ (2)

- wariancja

$$\eta_v(v) = \sum_{i=1}^M n_i |v_i - \bar{v}|^2 , \quad (3)$$

gdzie v_i jest wartością klasy kategorycznej i zaś \bar{v} jest średnią ze wszystkich obserwacji.

⁴[8]

Badanie zmienności parametru kontekstowego⁵ (3)

- unalikeability [5, 9]

$$\eta_u(v) = \frac{\sum_{i \neq j} c(x_i, x_j)}{n^2 - n} , \quad (4)$$

gdzie v jest zmienną kontekstową, x_i i x_j są obserwacjami zmiennej v , n jest liczbą wszystkich obserwacji zmiennej v , zaś $c(x_i, x_j) = 0$ jeżeli $x_i = x_j$ i $c(x_i, x_j) = 1$ jeżeli $x_i \neq x_j$.

⁵[8]

W jaki sposób interpretować wartość zmienności?

Procedura wyboru parametrów kontekstowych [8]

- Obliczamy zmienność każdego potencjalnego atrybutu kontekstowego na poziomie populacji.
- Obliczamy zmienność każdego potencjalnego atrybutu kontekstowego na poziomie użytkownika.
- Odrzucamy nieistotne na poziomie populacji atrybuty kontekstowe.

Przykład – wyznaczanie *unlikeability*

| User | Item (Movie) | Rating | Companion | Day |
|-------|-----------------------|--------|------------|----------|
| Alice | Donnie Darko | 1 | friend | Saturday |
| Alice | Girl Interrupted | 2 | friend | Friday |
| Alice | Shrek | 5 | family | Saturday |
| Alice | Spiderman | 1 | family | Sunday |
| Alice | The Counselor | 4 | friend | Friday |
| Alice | The Lion King | 4 | family | Sunday |
| Bob | An Unexpected Journey | 5 | alone | Saturday |
| Bob | City Of Angels | 2 | girlfriend | Saturday |
| Bob | Armageddon | 2 | alone | Friday |
| Bob | Inception | 1 | alone | Tuesday |
| Bob | Green Mile | 5 | alone | Saturday |
| Bob | Hunger Games | 2 | alone | Saturday |
| Bob | Tourist | 4 | girlfriend | Friday |
| Bob | Sleepless In Seattle | 4 | girlfriend | Friday |

- Przewidywana ocena użytkownika u dla przedmiotu i jest wyliczana ze wzoru:

$$\hat{r}_{ui} = \mu + b_i + b_u + \mathbf{q}_i^T \left(\mathbf{p}_u + |R(u)|^{\frac{1}{2}} \sum_{j \in R(u)} y_j \right), \quad (5)$$

gdzie μ jest średnią globalną ze wszystkich ocen w macierzy ocen, b_i i b_u są odpowiednio stronniczością produktu i użytkownika, $R(u)$ to produkty z historii użytkownika u , zaś wektor y_j jest reprezentacją produktu j z historii $R(u)$ w przestrzeni cech ukrytych.

$$\hat{r}_{ui} = \mu + b_i(t) + b_u(t) + \mathbf{q}_i^T \left(\mathbf{p}_u(t) + |R(u)|^{\frac{1}{2}} \sum_{j \in R(u)} y_j \right) , \quad (6)$$

gdzie t jest parametrem czasu.

$$\hat{r}_{ui} = \mathbf{r}_u^T \mathbf{w}_i \quad , \quad (7)$$

gdzie $r_{ui} = 0$.

$$\begin{aligned} \min_{\mathbf{w}_j} \quad & \frac{1}{2} \|\mathbf{r}_j - \mathbf{R}\mathbf{w}_j\|_2^2 + \frac{\beta}{2} \|\mathbf{w}_j\|_2^2 + \lambda \|\mathbf{w}_j\|_1 \\ \mathbf{w}_j \geqslant 0 \\ w_{jj} = 0 \end{aligned} \quad (8)$$

Algorytm CSLIM [12]

– reprezentacja kontekstu

- Założymy, że mamy dwa parametry kontekstowe: *czas* i *miejsce*.
- Parametr *czas* może przyjmować dwie wartości: "dzień roboczy" i "weekend".
- Parametr *miejsce* może przyjmować dwie wartości: "uczelnia" i "dom".
- Każdą sytuację kontekstową jesteśmy w stanie opisać wektorem [*czas*="dzień roboczy", *czas*="weekend", *miejsce*=uczelnia", *miejsce*="dom"].
- Sytuację kontekstową {"dzień roboczy", "uczelnia"} jesteśmy w stanie zapisać w postaci wektora binarnego $c = [1, 0, 1, 0]$.

Algorytm CSLIM [12]

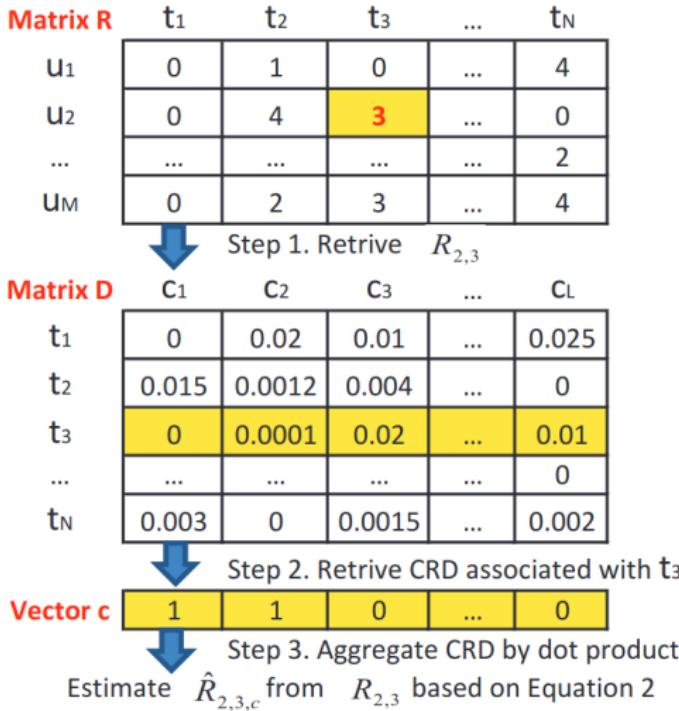
$$\hat{R}_{u,i,c} = R_{u,i} + \sum_{l=1}^L D_{i,l} c_l,$$

gdzie L to liczba warunków kontekstowych, $D_{i,l}$ jest macierzą odchyлеń, zaś c_l jest sytuacją kontekstową zapisaną w formacie binarnym.

$$\hat{S}_{u,i,c} = \sum_{h=1 \wedge h \neq i}^N (R_{u,h} + \sum_{l=1}^L D_{h,l} c_l) W_{h,i}$$

$$\begin{aligned} \min_{W,D} \quad & \frac{1}{2} \|R_{u,i,c} - \hat{S}_{u,i,c}\|_2^2 + \frac{\beta_1}{2} \|D\|_2^2 + \frac{\beta_2}{2} \|W\|_2^2 + \lambda_1 \|D\|_1 + \lambda_2 \|W\|_1 \\ \text{s.t.} \quad & W \geq 0 \\ & w_{jj} = 0 \quad . \end{aligned} \tag{9}$$

Algorytm CSLIM – macierz odchyleń



Rysunek: Zaczerpnięte z [12].

Jakiego rodzaju jest algorytm CSLIM?

- ang. *Context-Aware Splitting Approaches*
- Bazuje na pomyśle, że jeden produkt skonsumowany w różnych okolicznościach jest innym produktem.
- To samo może się odnosić również do użytkownika.
- Wyróżniamy trzy rodzaje podziałów:
 - względem produktu (ang. *Item Splitting*),
 - względem użytkownika (ang. *User Splitting*),
 - względem obu (ang. *User-Item Splitting*).

| Użytkownik | Miejsce | Ocena | Czas | Pogoda | Towarzystwo |
|------------|---------|-------|------------------|------------|-------------|
| u1 | p1 | 2 | Weekend | Pochmurnie | Dzieci |
| u1 | p1 | 4 | Weekend | Słonecznie | Rodzina |
| u1 | p1 | ? | Dzień roboczy | Deszczowo | Przyjaciel |

Tabela: Oceny miejsc w kontekście (na podstawie [11]).

Przykład – podział względem produktu

towarzystwo = *dzieci i nie dzieci*

| Użytkownik | Miejsce | Ocena |
|------------|---------|-------|
| u1 | p11 | 2 |
| u1 | p12 | 4 |
| u1 | p12 | ? |

Tabela: Macierz ocen po przekształceniu przez *item splitting* (na podstawie [11]).

pogoda = słonecznie i nie słonecznie

| Użytkownik | Miejsce | Ocena |
|------------|---------|-------|
| u11 | p1 | 2 |
| u12 | p1 | 4 |
| u11 | p1 | ? |

Tabela: Macierz ocen po przekształceniu przez *item splitting* (na podstawie [11]).

Przykład – podział względem obu

towarzystwo = dzieci i nie dzieci pogoda = słonecznie i nie słonecznie

| Użytkownik | Miejsce | Ocena |
|------------|---------|-------|
| u11 | p11 | 2 |
| u12 | p12 | 4 |
| u11 | p12 | ? |

Tabela: Macierz ocen po przekształceniu przez *user-item splitting* (na podstawie [11]).

Jakiego rodzaju jest metoda CASA?

W jaki sposób oceniać jakość kontekstowych systemów rekomendacyjnych?

Ocena jakości kontekstowych systemów rekomendacyjnych

- Analogicznie jak dla tradycyjnych dwuwymiarowych systemów.
- Agregujemy wyniki uzyskane z miar jakości per użytkownik i kontekst.
- Warto porównać wyniki z bazowym dwuwymiarowym algorytmem (o ile metoda kontekstowa korzysta z takiego).

Problemy oceny kontekstowych systemów rekomendacyjnych

- Użytkownicy powinni ocenić te same produkty w różnych kontekstach.
- Dostępne zbiory danych są bardzo małe i rzadkie, zwykle zbierane w ankietach.
- Trudności w porównaniu wyników.

Podsumowanie

Jakie problemy występują w systemach rekomedacyjnych?

- Problem "zimnego startu" (ang. *cold-start problem*)
 - Problem nowego użytkownika
 - Problem nowego produktu
 - Problem nowego systemu
- Rzadkość danych
- Nadmierne dopadowanie do danych
- Przewidywalność rekomendacji
- Zmiennaść preferencji użytkownika
- Złożoność obliczeniowa/skalowalność
- Problemy etyczne - prywatność użytkowników, np. RODO
- Zaufanie użytkowników

Dla jakich rodzajów systemów/algorytmów rekomendacji występują te problemy?

W jakich nie występują?

W jaki sposób zwiększyć zaufanie użytkowników do systemu rekomendacji?

Wyjaśnienia dla użytkowników

Rodzaje wyjaśnień

- funkcyjne,

Rodzaje wyjaśnień

- funkcyjne,
- przyczynowe,

Rodzaje wyjaśnień

- funkcyjne,
- przyczynowe,
- intencyjne,

Rodzaje wyjaśnień

- funkcyjne,
- przyczynowe,
- intencyjne,
- wyjaśnienia naukowe.

Czym są wyjaśnienia w systemach rekomendacyjnych?

Dodatkowa informacja prezentowana użytkownikowi w celu wyjaśnienia mu zwróconych wyników rekomendacji zgodnie z jakimś kryterium.

- transparentność,

- transparentność,
- weryfikowalność,

- transparentność,
- weryfikowalność,
- zaufanie,

Cele wyjaśnienia dla użytkownika

- transparentność,
- weryfikowalność,
- zaufanie,
- przekonywanie,

- transparentność,
- weryfikowalność,
- zaufanie,
- przekonywanie,
- skuteczność,

- transparentność,
- weryfikowalność,
- zaufanie,
- przekonywanie,
- skuteczność,
- satysfakcja,

- transparentność,
- weryfikowalność,
- zaufanie,
- przekonywanie,
- skuteczność,
- satysfakcja,
- zrozumiałość.

- G. Adomavicius and A. Tuzhilin. Handbook on recommender systems. chapter Context-Aware Recommender Systems, pages 217–256. Springer, 2011.
- M. Benerecetti, P. Bouquet, and C. Ghidini. Contextual reasoning distilled. *Philosophical Foundations of Artificial Intelligence. A special issue of the journal of Experimental and Theoretical AI (JETAI)*, 12(3):279–305, 2000.
- A. K. Dey. Understanding and using context. *Personal Ubiquitous Comput.*, 5(1):4–7, Jan. 2001.
- A. K. Dey and G. D. Abowd. Towards a better understanding of context and context-awareness. In *Proc. of Conference on Human Factors in Computing Systems*, pages 304–307, 2000.
- G. D. Kader and M. Perry. Variability for categorical variables. *Journal of Statistics Education*, 15(2), 2007.
- A. Karpus. Context-Aware User Modelling and Generation of Recommendations in Recommender Systems. Rozprawa doktorska. 2017.
- Y. Koren. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*, pages 447–456, New York, NY, USA, 2009. ACM.
- A. Odic, M. Tkalcic, J. F. Tasic, and A. Kosir. Predicting and detecting the relevant contextual information in a movie-recommender system. *Interacting with Computers*, 25(1):74–90, 2013.
- M. Perry and G. D. Kader. Variation as unalikeability. *Teaching Statistics*, 27(2):58–60, 2005.
- L. Serafini and P. Bouquet. Comparing formal theories of context in ai. *Artif. Intell.*, 155(1-2):41–67, May 2004.
- Y. Zheng, R. Burke, and B. Mobasher. Splitting approaches for context-aware recommendation: An empirical study. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing, SAC '14*, pages 274–279, New York, NY, USA, 2014. ACM.
- Y. Zheng, B. Mobasher, and R. Burke. Cslim: Contextual slim recommendation algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14*, pages 301–304, New York, NY, USA, 2014. ACM.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Analiza szeregów czasowych

dr inż. Aleksandra Karpus

10 grudnia 2021



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Czym jest szereg czasowy?

Szereg czasowy

Szereg czasowy (ang. *Time-Series Data, Time-related data*)

to ciąg obserwacji pewnego zjawiska w kolejnych jednostkach czasu.

Szereg czasowy (ang. *Time-Series Data, Time-related data*)

to ciąg obserwacji pewnego zjawiska w kolejnych jednostkach czasu.

Założenia:

- odstępy czasu są równe,
- dane są uporządkowane chronologicznie.

Szereg czasowy

Szereg czasowy (ang. *Time-Series Data, Time-related data*)

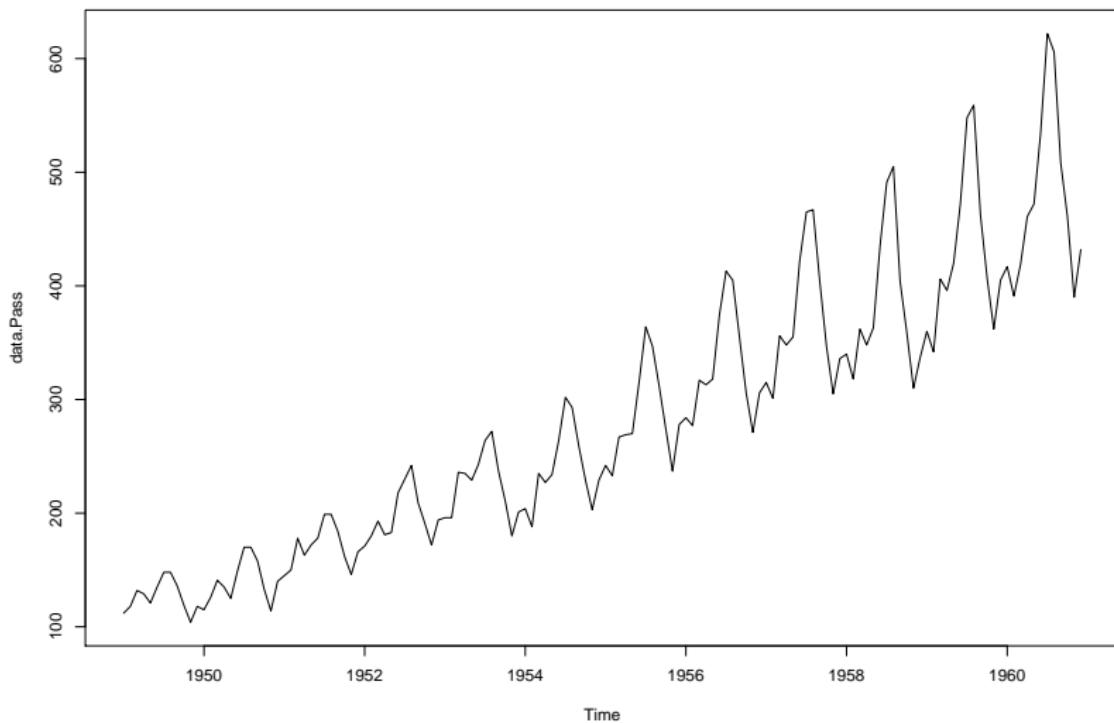
to ciąg obserwacji pewnego zjawiska w kolejnych jednostkach czasu.

Założenia:

- odstępy czasu są równe,
- dane są uporządkowane chronologicznie.

Przedział czasu, w którym zbierane są dane, jest ogólnie określany jako **częstotliwość szeregów czasowych**.

Przykład szeregu czasowego



Typy szeregów czasowych

Szereg czasowy momentów

szereg zawierający informacje o poziomach badanego zjawiska w określonych momentach pewnego przedziału czasowego.

Szereg czasowy momentów

szereg zawierający informacje o poziomach badanego zjawiska w określonych momentach pewnego przedziału czasowego.

Szereg czasowy okresów

szereg zawierający informacje o rozmiarach zjawiska w ciągu kolejnych okresów danego przedziału czasowego.

Czym różni się szereg czasowy od sekwencji danych?

1. Określenie celu analizy
2. Poznanie danych
3. Analiza podstawowa/wstępna
4. Identyfikacja obserwacji odstających i brakujących
5. Rozłożenie szeregu czasowego na czynniki pierwsze
6. Analiza właściwa (zgodna z celem)
7. Ocena wyników i oszacowanie ryzyka

Które kroki analizy szeregów czasowych pokrywają się z procesem eksploracji danych?

Jaki jest cel analizy?

- Badanie występujących w danych: regularnych cykli, wzorców, trendów
- Prognozowanie wartości szeregów dla przyszłych okresów, na podstawie obserwacji historycznych
- Znalezienie modelu dobrze opisującego przebieg danego zjawiska w czasie
- Porównanie zjawiska z dwóch obserwacji – poszukiwanie wzorców, podobieństw
- Poszukiwanie/wykrywanie anomalii

Przykładowe cele:

- Poszukiwanie wzorców sprzedażowych.
- Przewidywanie wartości sprzedaży w kolejnym okresie.

Szereg czasowy sprzedaży odzieży męskiej:

[http:](http://smartdrill.com/images/sales%20of%20men's%20clothing.jpg)

[//smartdrill.com/images/sales%20of%20men's%20clothing.jpg](http://smartdrill.com/images/sales%20of%20men's%20clothing.jpg)

Przykładowe cele:

- Przewidywanie wartości akcji w kolejnym okresie.
- Szukanie modelu zmian na giełdzie.

Szereg czasowy cen akcji IBM i LinkedIn na giełdzie:

<https://revolution-computing.typepad.com/.a/6a010534b1db25970b01b7c7c4a75a970b-800wi>

Przykładowe cele:

- Porównanie zjawiska z dwóch obserwacji.

Szeregi czasowe danych pomiarowych z dwóch mierników:

https://www.mdpi.com/remotesensing/remotesensing-08-00970/article_deploy/html/images/remotesensing-08-00970-g001-550.jpg

Przykładowe cele:

- Wykrywanie anomalii w przebiegu.
- Badanie wzorców i trendów.

Szeregi czasowe danych sprzedażowych różnych produktów w USA:

https://people.duke.edu/~rnau/411infla_files/image002.png

Przykładowe cele:

- Znalezienie zależności pomiędzy zjawiskami.
- Wyjaśnienie anomalii/przebiegu szeregu czasowego.

Szeregi czasowe danych z parku Yellowstone:

[https://cdn.aptech.com/www/uploads/2019/09/
ts-pp-yellowstone-1.jpg](https://cdn.aptech.com/www/uploads/2019/09/ts-pp-yellowstone-1.jpg)

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?
 - Uwaga na metryki stosunkowe i procentowe! Uwaga na jednostki pieniężne w latach! Uwaga na pomiar zmiany!
 - Jaka jest dokładność pomiaru?

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?
 - Uwaga na metryki stosunkowe i procentowe! Uwaga na jednostki pieniężne w latach! Uwaga na pomiar zmiany!
 - Jaka jest dokładność pomiaru?
- Czy są to dane „surowe” czy poddane już przetworzeniu? Jak były przetwarzane?

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?
 - Uwaga na metryki stosunkowe i procentowe! Uwaga na jednostki pieniężne w latach! Uwaga na pomiar zmiany!
 - Jaka jest dokładność pomiaru?
- Czy są to dane „surowe” czy poddane już przetworzeniu? Jak były przetwarzane?
- Czy dane są „czyste” czy „brudne”?

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?
 - Uwaga na metryki stosunkowe i procentowe! Uwaga na jednostki pieniężne w latach! Uwaga na pomiar zmiany!
 - Jaka jest dokładność pomiaru?
- Czy są to dane „surowe” czy poddane już przetworzeniu? Jak były przetwarzane?
- Czy dane są „czyste” czy „brudne”?
 - Czy dane zawierają błędy? Brakujące pomiary? Były mierzone w różnych okresach czasowych? Były zmiany procedury pomiarowej?

Poznanie danych

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?
 - Uwaga na metryki stosunkowe i procentowe! Uwaga na jednostki pieniężne w latach! Uwaga na pomiar zmiany!
 - Jaka jest dokładność pomiaru?
- Czy są to dane „surowe” czy poddane już przetworzeniu? Jak były przetwarzane?
- Czy dane są „czyste” czy „brudne”?
 - Czy dane zawierają błędy? Brakujące pomiary? Były mierzone w różnych okresach czasowych? Były zmiany procedury pomiarowej?
- Zobacz dane – wygeneruj wykres (punktowy, liniowy)

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach? Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?
 - Uwaga na metryki stosunkowe i procentowe! Uwaga na jednostki pieniężne w latach! Uwaga na pomiar zmiany!
 - Jaka jest dokładność pomiaru?
- Czy są to dane „surowe” czy poddane już przetworzeniu? Jak były przetwarzane?
- Czy dane są „czyste” czy „brudne”?
 - Czy dane zawierają błędy? Brakujące pomiary? Były mierzone w różnych okresach czasowych? Były zmiany procedury pomiarowej?
- Zobacz dane – wygeneruj wykres (punktowy, liniowy)
 - Czy wszystko się zgadza?
 - Czego można się spodziewać?

Analiza podstawowa

Analiza składowej stałej

- Analiza tendencji centralnej

- Średnia arytmetyczna
 - Średnia chronologiczna

$$\bar{y} = \frac{\frac{y_1+y_2}{2} + \frac{y_2+y_3}{2} + \dots + \frac{y_{n-1}+y_n}{2}}{n}$$

- Mediana

Analiza podstawowa

Analiza składowej stałej

- Analiza tendencji centralnej

- Średnia arytmetyczna
 - Średnia chronologiczna

$$\bar{y} = \frac{\frac{y_1+y_2}{2} + \frac{y_2+y_3}{2} + \dots + \frac{y_{n-1}+y_n}{2}}{n}$$

- Mediana

- Analiza zmienności

- wariancja, odchylenie standardowe
 - wartości minimalne i maksymalne, rozrzut
 - kwartyle, odstęp Q3-Q1

- Przyrosty
 - Przyrost absolutny jednopodstawowy (j - okres bazowy)

$$\Delta x_{n/j} = x_n - x_j$$

- Przyrost absolutny łańcuchowy

$$\Delta x_{n/n-1} = x_n - x_{n-1}$$

- Przyrost względny jednopodstawowy (j - okres bazowy)

$$\Delta x_{n/j} = \frac{x_n - x_j}{x_j}$$

- Przyrost względny łańcuchowy

$$\Delta x_{n/n-1} = \frac{x_n - x_{n-1}}{x_{n-1}}$$

- Indeksy dynamiki

- Indeks jednopodstawowy (j - okres bazowy)

$$x_{n/j} = \frac{x_n}{x_j}$$

- Indeks łańcuchowy

$$x_{n/n-1} = \frac{x_n}{x_{n-1}}$$

- Indeksy dynamiki

- Indeks jednopodstawowy (j - okres bazowy)

$$x_{n/j} = \frac{x_n}{x_j}$$

- Indeks łańcuchowy

$$x_{n/n-1} = \frac{x_n}{x_{n-1}}$$

- Inne miary

- średniookresowe tempo zmian
 - indeksy agregatowe

Przykład – przyrosty

| lata | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|--------------------------------|-------|-------|-------|-------|-------|-------|-------|
| urodzenia żywe w Polsce w tys. | 378,3 | 368,2 | 353,8 | 351,1 | 356,1 | 364,4 | 374,2 |

Tabela: Źródło: Roczniki demograficzne.

| Okresy czasu (lata) | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|--|------|------|------|------|------|------|------|
| Przyrosty absolutne w tys. (rok bazowy - 2000) | 0 | - | - | - | - | - | -4,1 |
| Przyrosty względne w % (rok bazowy - 2000) | 0 | -2,7 | -6,5 | -7,2 | -5,9 | -3,7 | -1,1 |
| Przyrosty absolutne w tys. (rok bazowy - 2003) | 27,2 | 17,1 | 2,7 | 0 | 5 | 13,3 | 23,1 |
| Przyrosty względne w % (rok bazowy - 2003) | 7,8 | 4,9 | 0,8 | 0 | 1,4 | 3,8 | 6,6 |

Tabela: Źródło przykładu:

http://www.demografia.uni.lodz.pl/dlastud/szeregi_czasowe_I.pdf

Przykład – indeksy

| lata | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|--------------------------------|-------|-------|-------|-------|-------|-------|-------|
| urodzenia żywe w Polsce w tys. | 378,3 | 368,2 | 353,8 | 351,1 | 356,1 | 364,4 | 374,2 |

Tabela: Źródło: Roczniki demograficzne.

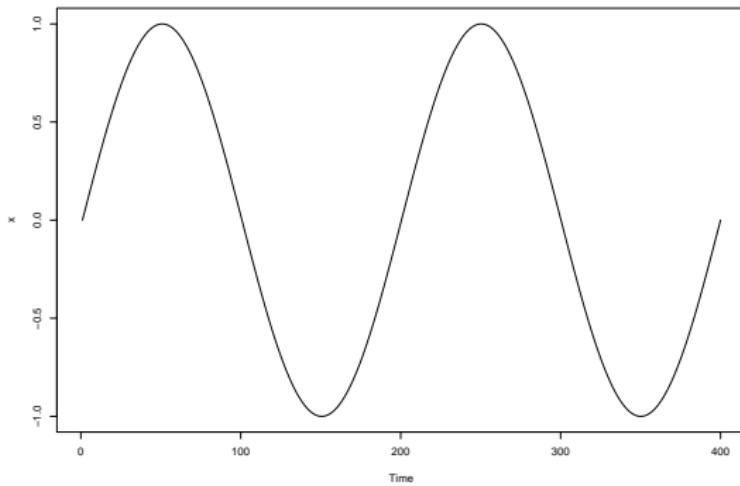
| Lata | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|--|------|------|------|------|-------|-------|-------|
| Indeksy łańcuchowe w % (rok poprzedni = 100) | - | 97,3 | 96,1 | 99,2 | 101,4 | 102,3 | 102,7 |
| Indeksy jednopodstawowe w % (rok 2000 = 100) | 100 | 97,3 | 93,5 | 92,8 | 94,1 | 96,3 | 98,9 |

Tabela: Źródło przykładu:

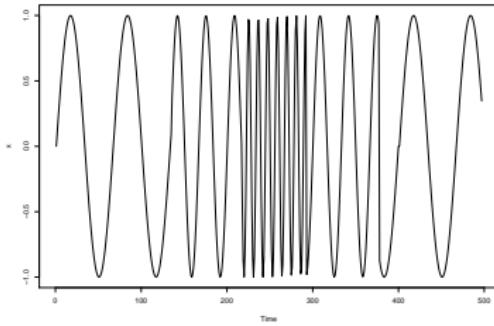
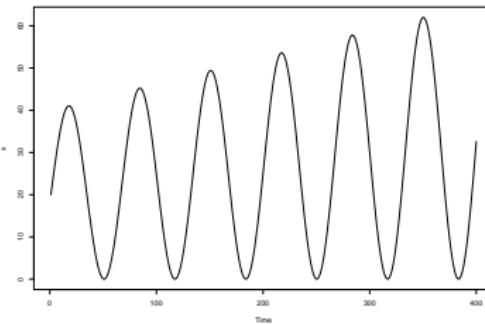
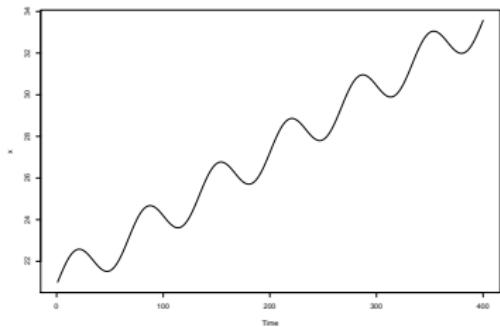
http://www.demografia.uni.lodz.pl/dlastud/szeregi_czasowe_I.pdf

Stacjonarność szeregu czasowego

Szereg czasowy, którego charakterystyki statystyczne (średnia, wariancja, autokorelacja) nie są zmienne w czasie, nazywamy **stacjonarnym**.



Przykłady szeregów niestacjonarnych



Po co stacjonaryzować szereg czasowy?

Przyczyny stacjonaryzowania szeregów czasowych

- Wiele metod analizy szeregów czasowych wymaga, aby był on stacjonarny.

Przyczyny stacjonaryzowania szeregow czasowych

- Wiele metod analizy szeregow czasowych wymaga, aby był on stacjonarny.
- Model zbudowany na szeregu niestacjonarnym będzie obarczony dużym błędem.

W jaki sposób ustacjonaryzować szereg czasowy?

Metody stacjonaryzowania szeregów czasowych

- Transformacja logarytmiczna

Metody stacjonaryzowania szeregów czasowych

- Transformacja logarytmiczna
- Pierwiastek kwadratowy

Metody stacjonaryzowania szeregów czasowych

- Transformacja logarytmiczna
- Pierwiastek kwadratowy
- Indeksy zmian

Metody stacjonaryzowania szeregów czasowych

- Transformacja logarytmiczna
- Pierwiastek kwadratowy
- Indeksy zmian
- Estymacja i odjęcie trendu

Metody stacjonaryzowania szeregów czasowych

- Transformacja logarytmiczna
- Pierwiastek kwadratowy
- Indeksy zmian
- Estymacja i odjęcie trendu
- Estymacja i odjęcie sezonowości

Metody stacjonaryzowania szeregów czasowych

- Transformacja logarytmiczna
- Pierwiastek kwadratowy
- Indeksy zmian
- Estymacja i odjęcie trendu
- Estymacja i odjęcie sezonowości

Jeśli po dwóch/trzech operacjach różnicowania szereg nadal jest niestacjonarny, to jego ustacjonaryzowanie w kolejnych iteracjach jest bardzo mało prawdopodobne.

Dekompozycja szeregu czasowego

- Część systematyczna
- Część przypadkowa – szum

- Część systematyczna
 - Składowa stała
- Część przypadkowa – szum

- Część systematyczna
 - Składowa stała
 - Trend
- Część przypadkowa – szum

- Część systematyczna
 - Składowa stała
 - Trend
 - Składowa okresowa
- Część przypadkowa – szum

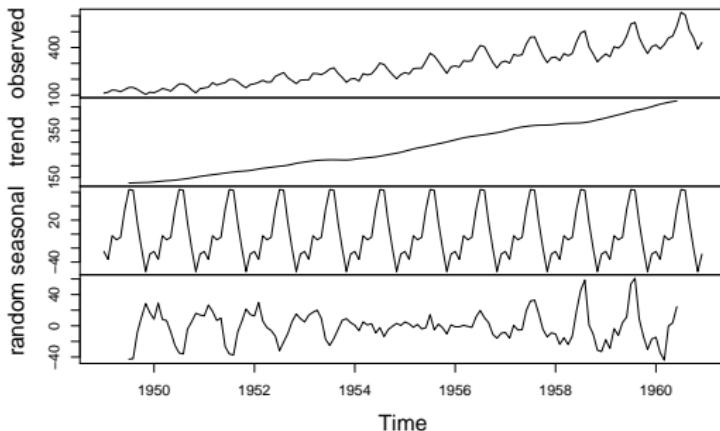
- Część systematyczna
 - Składowa stała
 - Trend
 - Składowa okresowa
 - wahania cykliczne – długookresowe rytmiczne wahania,
- Część przypadkowa – szum

- Część systematyczna
 - Składowa stała
 - Trend
 - Składowa okresowa
 - wahania cykliczne – długookresowe rytmiczne wahania,
 - wahania sezonowe – wahania wynikające "z kalendarza" (pory roku, itp.)
- Część przypadkowa – szum

- Część systematyczna
 - Składowa stała
 - Trend
 - Składowa okresowa
 - wahania cykliczne – długookresowe rytmiczne wahania,
 - wahania sezonowe – wahania wynikające "z kalendarza" (pory roku, itp.)
 - okresowość – krótkoterminowa powtarzalność zmiany amplitudy wynikająca z charakteru badanego zjawiska
- Część przypadkowa – szum

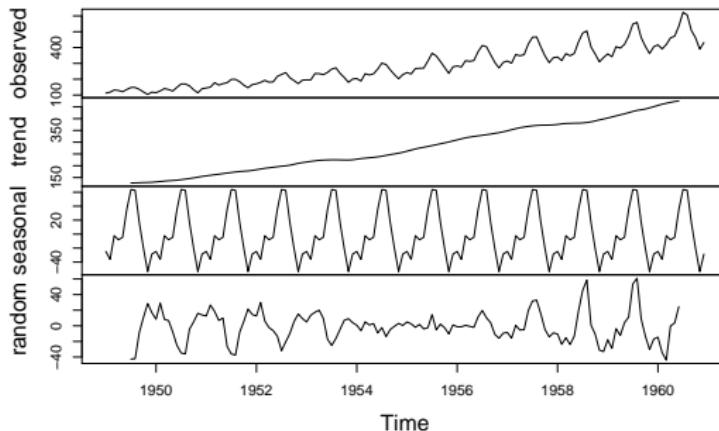
Dekompozycja szeregu czasowego

Decomposition of additive time series



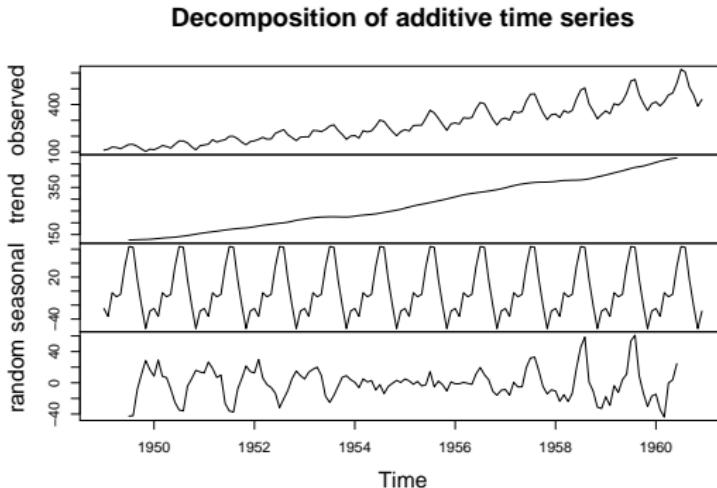
Dekompozycja szeregu czasowego

Decomposition of additive time series



Dekompozycję powtarzamy tak długo, aż pozostała część szeregu będzie szumem białym:

Dekompozycja szeregu czasowego



Dekompozycję powtarzamy tak długo, aż pozostała część szeregu będzie szumem białym:

- średnia = 0,
- wariancja stała w czasie,
- brak autokorelacji.

Po co wykonywać dekompozycję szeregu czasowego?

Cele dekompozycji szeregu czasowego

- Analiza trendu.
- Analiza cykliczności.
- Dopasowanie modelu.

Rodzaje modeli

- Addytywne

Dane = Trend + Sezonowość + Cykliczność + Czynnik losowy

Rodzaje modeli

- Addytywne

Dane = Trend + Sezonowość + Cykliczność + Czynnik losowy

- Multiplikatywne

Dane = Trend * Sezonowość * Cykliczność * Czynnik losowy

- Addytywne

Dane = Trend + Sezonowość + Cykliczność + Czynnik losowy

- Mnożnikowe

Dane = Trend * Sezonowość * Cykliczność * Czynnik losowy

- Mieszane.

- Addytywne

Dane = Trend + Sezonowość + Cykliczność + Czynnik losowy

- Multiplikatywne

Dane = Trend * Sezonowość * Cykliczność * Czynnik losowy

- Mieszane.

Graficzne porównanie modeli: https://miro.medium.com/max/1400/1*rDQL2fAp_X_dgAHNZuwRfw.png

1. Agnieszka Landowska, *Materiały wykładowe do przedmiotu Przetwarzanie Danych w Biznesie*, Politechnika Gdańsk, 2020.
2. Aileen Nielsen, *Szeregi czasowe. Praktyczna analiza i predykcja z wykorzystaniem statystyki i uczenia maszynowego*, Helion, 2020.
3. Ane Blázquez-Garcia, Angel Conde, Usue Mori and José Antonio Lozano, *A Review on Outlier/Anomaly Detection in Time Series Data*. ACM Computing Surveys (CSUR) 54 (2021): 1 - 33.
4. Avril Coghlan, *A Little Book of R For Time Series*, 2016,
<https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf>.
5. Robert Nau, *Principles and risks of forecasting*, Fuqua School of Business, Duke University, 2014,
https://people.duke.edu/~rnau/Principles_and_risks_of_forecasting--Robert_Nau.pdf.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Analiza szeregów czasowych

dr inż. Aleksandra Karpus

17 grudnia 2021



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



KANCELARIA PREZESA RADY MINISTRÓW
MINISTERSTWO ROZWOJU REGIONALNEGO

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Autokorelacja

Autokorelacja to związek pomiędzy sygnałem a jego opóźnioną kopią w funkcji opóźnienia.

Autokorelacja

Autokorelacja to związek pomiędzy sygnałem a jego opóźnioną kopią w funkcji opóźnienia.

Autokorelacja daje nam wyobrażenie o liniowym powiązaniu pomiędzy danymi znajdującymi się w szeregu w zależności od odległości pomiędzy nimi..

Autokorelacja to związek pomiędzy sygnałem a jego opóźnioną kopią w funkcji opóźnienia.

Autokorelacja daje nam wyobrażenie o liniowym powiązaniu pomiędzy danymi znajdującymi się w szeregu w zależności od odległości pomiędzy nimi..

Autokorelację możemy badać za pomocą dwóch funkcji:

Autokorelacja to związek pomiędzy sygnałem a jego opóźnioną kopią w funkcji opóźnienia.

Autokorelacja daje nam wyobrażenie o liniowym powiązaniu pomiędzy danymi znajdującymi się w szeregu w zależności od odległości pomiędzy nimi..

Autokorelację możemy badać za pomocą dwóch funkcji:

- funkcja autokorelacji (ang. *autocorrelation function*, ACF),

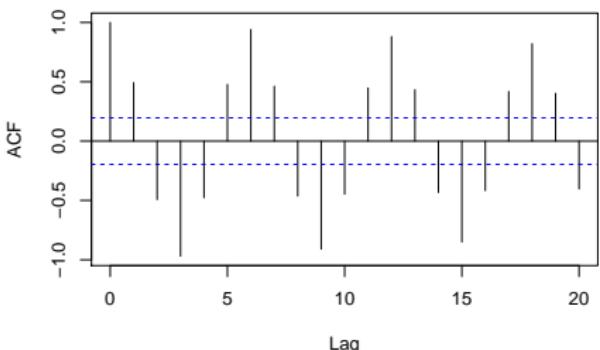
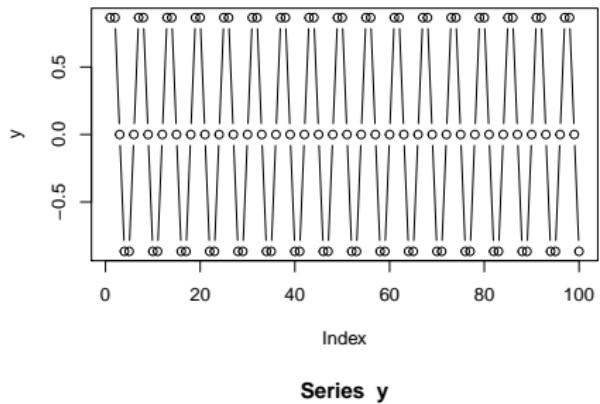
Autokorelacja to związek pomiędzy sygnałem a jego opóźnioną kopią w funkcji opóźnienia.

Autokorelacja daje nam wyobrażenie o liniowym powiązaniu pomiędzy danymi znajdującymi się w szeregu w zależności od odległości pomiędzy nimi..

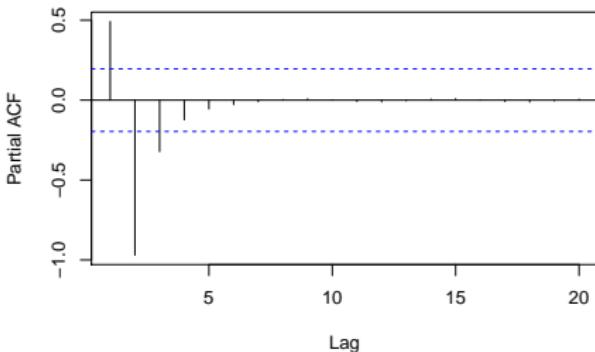
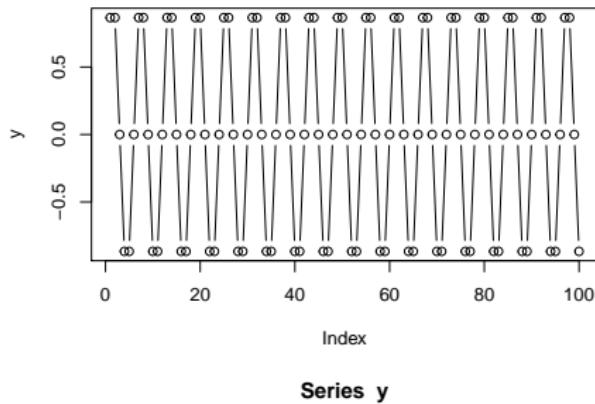
Autokorelację możemy badać za pomocą dwóch funkcji:

- funkcja autokorelacji (ang. *autocorrelation function*, ACF),
- funkcja autokorelacji cząstkowej (ang. *partial autocorrelation function*, PACF).

Funkcja autokorelacji



Funkcja autokorelacji cząstkowej



Do analizy szeregu czasowego oraz predykcji kolejnych okresów wykorzystywane są modele statystyczne:

- Modele autoregresyjne (AR),
- Modele ze średnią ruchomą (MA),
- Zintegrowane modele autoregresyjne średniej ruchomej (ARMA/ARIMA),
- Inne modele.

Dlaczego nie należy wykorzystywać regresji liniowej do predykcji kolejnych wartości szeregu czasowego?

Ograniczenia regresji liniowej przy analizie szeregów czasowych

- Regresja liniowa zakłada niezależność analizowanych obserwacji,

Ograniczenia regresji liniowej przy analizie szeregów czasowych

- Regresja liniowa zakłada niezależność analizowanych obserwacji,
- Regresja zakłada, że dane mają jednakowy rozkład,

Ograniczenia regresji liniowej przy analizie szeregów czasowych

- Regresja liniowa zakłada niezależność analizowanych obserwacji,
- Regresja zakłada, że dane mają jednakowy rozkład,
- Założenia te nie są spełnione dla szeregów czasowych.

- Opiera się na intuicji, że przeszłość przewiduje przyszłość.

- Opiera się na intuicji, że przeszłość przewiduje przyszłość.
- Zakłada, że w danym szeregu wartość w momencie t jest funkcją wartości we wcześniejszych punktach czasu.

- Opiera się na intuicji, że przeszłość przewiduje przyszłość.
- Zakłada, że w danym szeregu wartość w momencie t jest funkcją wartości we wcześniejszych punktach czasu.
- Najprostszy model autoregresji AR(1) ma postać

$$y_t = b_0 + b_1 y_{t-1} + e_t,$$

gdzie b_0 i b_1 to stałe, zaś e_t jest zmiennym w czasie błędem ze stałą wariancją i zerową średnią.

- Główną ideą tego modelu jest koncepcja wielu niezależnych zdarzeń zachodzących w różnych momentach czasu i wpływających indywidualnie na bieżącą wartość szeregu.

- Główną ideą tego modelu jest koncepcja wielu niezależnych zdarzeń zachodzących w różnych momentach czasu i wpływających indywidualnie na bieżącą wartość szeregu.
- Zakłada się, że wartość szeregu w danym punkcie czasu jest opisywana przez funkcję ostatnich wartości składników "błędu", które są niezależne od siebie.

- Główną ideą tego modelu jest koncepcja wielu niezależnych zdarzeń zachodzących w różnych momentach czasu i wpływających indywidualnie na bieżącą wartość szeregu.
- Zakłada się, że wartość szeregu w danym punkcie czasu jest opisywana przez funkcję ostatnich wartości składników "błędu", które są niezależne od siebie.
- Model MA(q) na postać:

$$y_t = \mu + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots + \theta_q e_{t-q},$$

gdzie $\mu, \theta_1, \dots, \theta_q$ to stałe.

Zintegrowane modele autoregresyjne średniej ruchomej (ARMA/ARIMA)

- ARMA jest połączeniem modeli autoregresyjnego i średniej ruchomej.

Zintegrowane modele autoregresyjne średniej ruchomej (ARMA/ARIMA)

- ARMA jest połączeniem modeli autoregresyjnego i średniej ruchomej.
- Model ARMA ma postać:

$$y_t = \phi_0 + \sum(\phi_i r_{t-i}) + e_t - \sum(\theta_i e_{t-i}).$$

Zintegrowane modele autoregresyjne średniej ruchomej (ARMA/ARIMA)

- ARMA jest połączeniem modeli autoregresyjnego i średniej ruchomej.
- Model ARMA ma postać:

$$y_t = \phi_0 + \sum(\phi_i r_{t-i}) + e_t - \sum(\theta_i e_{t-i}).$$

- Model ARIMA wykorzystuje dodatkowo operację różnicowania.

Zintegrowane modele autoregresyjne średniej ruchomej (ARMA/ARIMA)

- ARMA jest połączeniem modeli autoregresyjnego i średniej ruchomej.
- Model ARMA ma postać:

$$y_t = \phi_0 + \sum(\phi_i r_{t-i}) + e_t - \sum(\theta_i e_{t-i}).$$

- Model ARIMA wykorzystuje dodatkowo operację różnicowania.
- Z twierdzenia Wolda wynika, że dla każdego szeregu stacjonarnego istnieje dobrze przybliżający go model ARMA,

Zintegrowane modele autoregresyjne średniej ruchomej (ARMA/ARIMA)

- ARMA jest połączeniem modeli autoregresyjnego i średniej ruchomej.
- Model ARMA ma postać:

$$y_t = \phi_0 + \sum(\phi_i r_{t-i}) + e_t - \sum(\theta_i e_{t-i}).$$

- Model ARIMA wykorzystuje dodatkowo operację różnicowania.
- Z twierdzenia Wolda wynika, że dla każdego szeregu stacjonarnego istnieje dobrze przybliżający go model ARMA,
- ... choć jego znalezienie może nie być łatwe.

W jaki sposób dokonać wyboru modelu najlepiej opisującego dany szereg czasowy?

| Rodzaj wykresu | ACF | PACF |
|----------------|---|---|
| AR(p) | powoli malejący | ostry spadek po minieciu opóźnienia p |
| MA(q) | ostry spadek po minieciu opóźnienia q | powoli malejący |
| ARMA | brak wyraźnego punktu odcięcia | brak wyraźnego punktu odcięcia |

Jakie są zalety stosowania omówionych modeli statystycznych?

- Prostota modeli.
- Dają dobre wyniki już dla małych zbiorów danych.
- Zwracają prognozy dobrej jakości bez ryzyka nadmiernego dopasowania, nawet w porównaniu ze skomplikowanymi metodami uczenia maszynowego.
- Pozwalają na szybkie tworzenie prognoz dzięki dobrym i automatycznym metodom doboru parametrów modeli.

A jakie są ich wady?

- Nie są zbyt wydajne na dużych zbiorach danych.
- Metody te szacują średnią punktową wartość z rozkładu, a nie sam rozkład, przez co są obarczone dużą niepewnością.
- Nie są przystosowane do obsługi dynamiki nieliniowej. Zastosowanie ich na danych, w których występują nieliniowe relacje, da słabe rezultaty.

Ocena dopasowania modelu

- Ocena pozostałości po odjęciu modelu od danych szeregu czasowego.
 - Bliskość do szumu białego.

Ocena dopasowania modelu

- Ocena pozostałości po odjęciu modelu od danych szeregu czasowego.
 - Bliskość do szumu białego.
- Minimalizacja błędu prognozy:
 - Błąd średniokwadratowy,
 - Pierwiastek z błędu średniokwadratowego,
 - Znormalizowany błąd średniokwadratowy

$$NMSE = \frac{\sum_{j=1}^N (\text{obserwacja}_j - \text{predykcja}_j)^2}{\sum_{j=1}^N (\text{obserwacja}_j - \text{średnia})^2},$$

- Średni bezwględny błąd procentowy

$$MAPE = \frac{100}{N} \sum_{i=0}^N \frac{|\text{prognoza}_i - \text{prawdziwa}_i|}{\text{prawdziwa}_i}.$$

- Ryzyko wewnętrzne

- występuje zawsze,
 - można oszacować standardowy błąd,
 - redukcja przez poszukiwanie wzorców – trendów, sezonowości itp.

- Ryzyko wewnętrzne
 - występuje zawsze,
 - można oszacować standardowy błąd,
 - redukcja przez poszukiwanie wzorców – trendów, sezonowości itp.
- Ryzyko doboru parametrów
 - redukcja przez zastosowanie lepiej dopasowanych metod,
 - więcej danych może nie pomóc.

Rodzaje ryzyka predykcji

- Ryzyko wewnętrzne
 - występuje zawsze,
 - można oszacować standardowy błąd,
 - redukcja przez poszukiwanie wzorców – trendów, sezonowości itp.
- Ryzyko doboru parametrów
 - redukcja przez zastosowanie lepiej dopasowanych metod,
 - więcej danych może nie pomóc.
- Ryzyko wyboru modelu
 - redukcja przez analizę danych, sprawdzenie założeń itp.
 - żadna analiza nie dostarczy oszacowania takiego błędu,
 - mierzalny skutkami.

Inne cele analizy szeregu czasowego

- Porównywanie szeregów.

Inne cele analizy szeregu czasowego

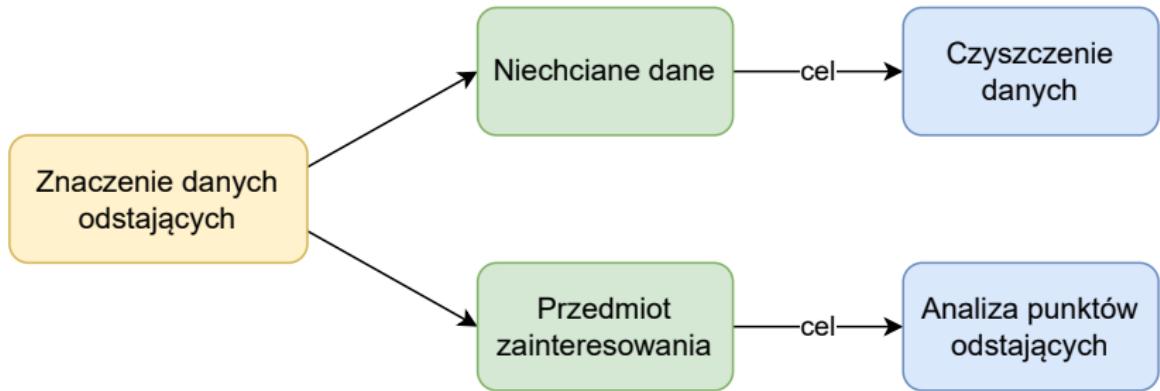
- Porównywanie szeregow.
- Wykrywanie motywów.

Inne cele analizy szeregu czasowego

- Porównywanie szeregow.
- Wykrywanie motywów.
- **Wykrywanie anomalii.**

Anomalie to punkty danych, które znacznie odbiegają od ogólnego zachowania danych.

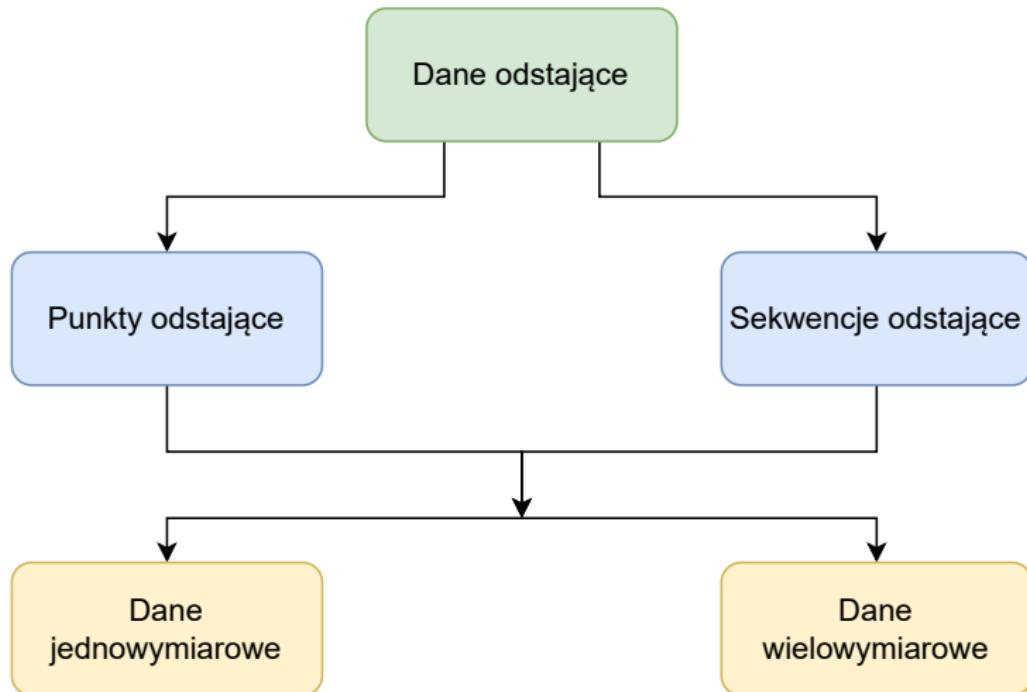
Anomalie a cel ich wykrywania



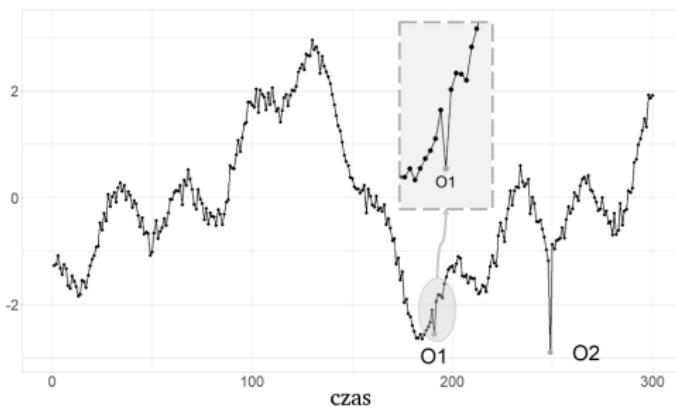
Po co analizować anomalie?

- Wykrywanie oszustw finansowych
- Filtrowanie spamu
- Identyfikacja wadliwych urządzeń/czujników
- Wykrywanie anomalii użycia procesora
- i inne

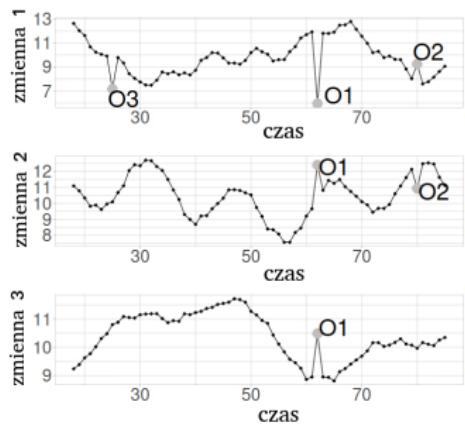
Rodzaje anomalii



Punkt odstający



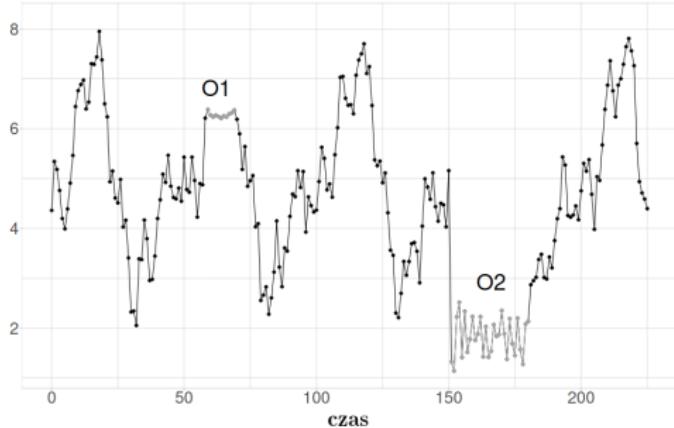
(a) Jednowymiarowy szereg czasowy



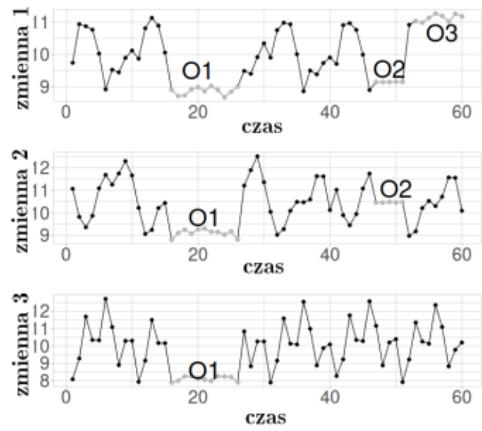
(b) Wielowymiarowy szereg czasowy

Rysunek: Wykres zaczerpnięty z [3].

Sekwencja odstająca



(a) Jednowymiarowy szereg czasowy



(b) Wielowymiarowy szereg czasowy

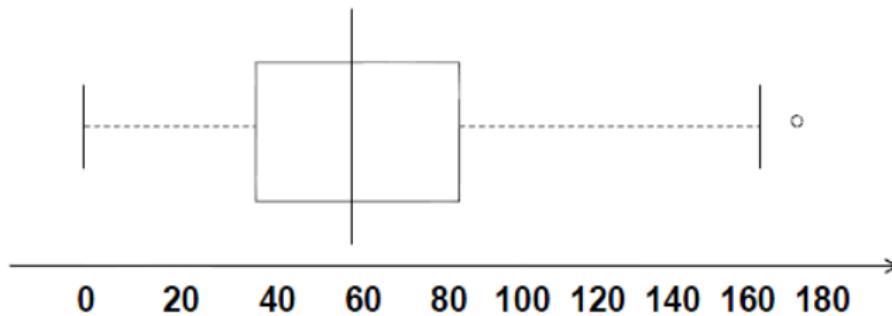
Rysunek: Wykres zaczerpnięty z [3].

W jaki sposób wykrywać anomalie?

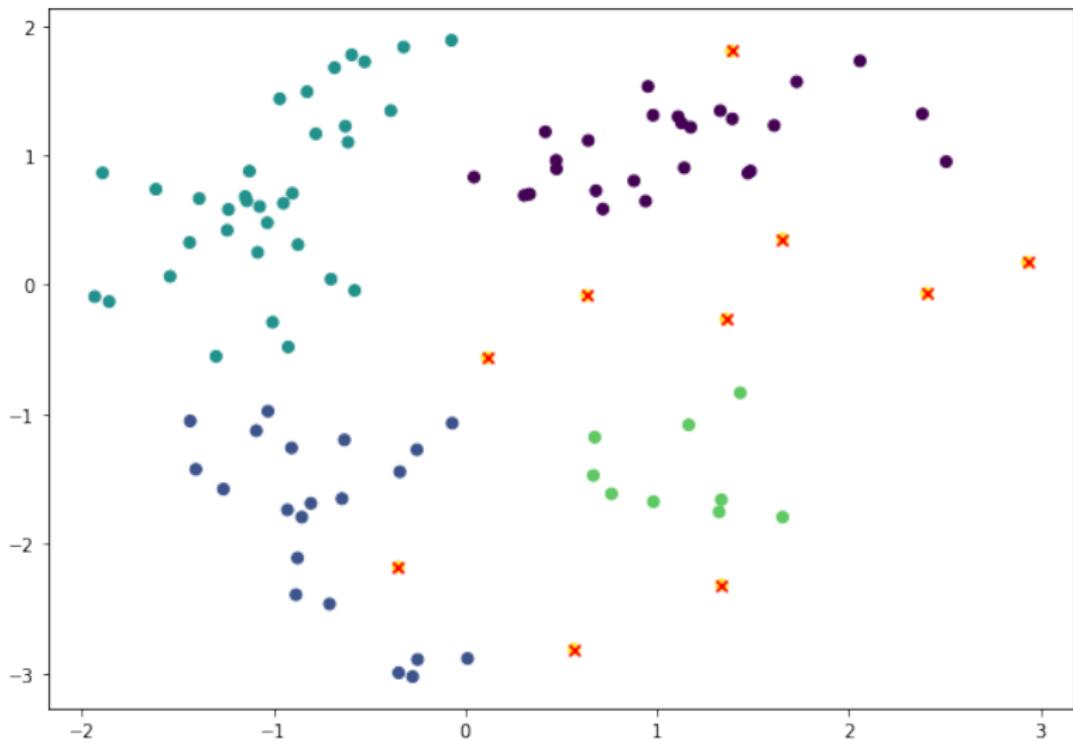
Metody wykrywania anomalii w szeregach czasowych

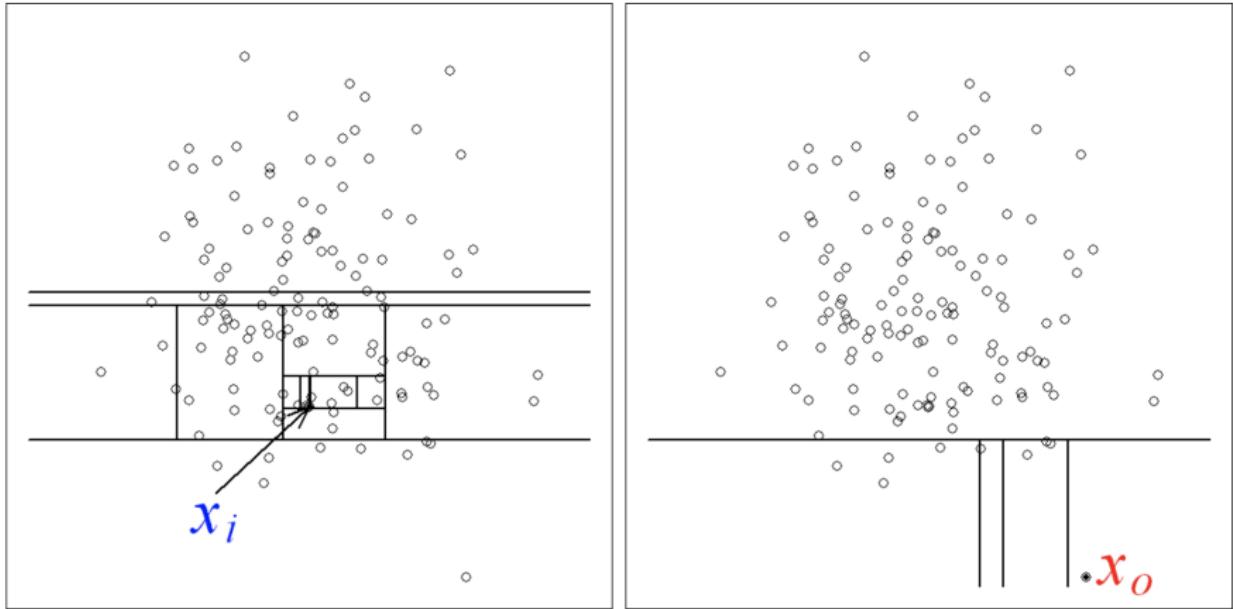
- IQR/wykresy pudełkowe
- Algorytm K-średnich
- Las izolacji (ang. *Isolation forest*)
- Dekompozycja szeregu czasowego
- Wykorzystanie modelu predykcyjnego
- Autoenkodery

IQR/wykresy pudełkowe



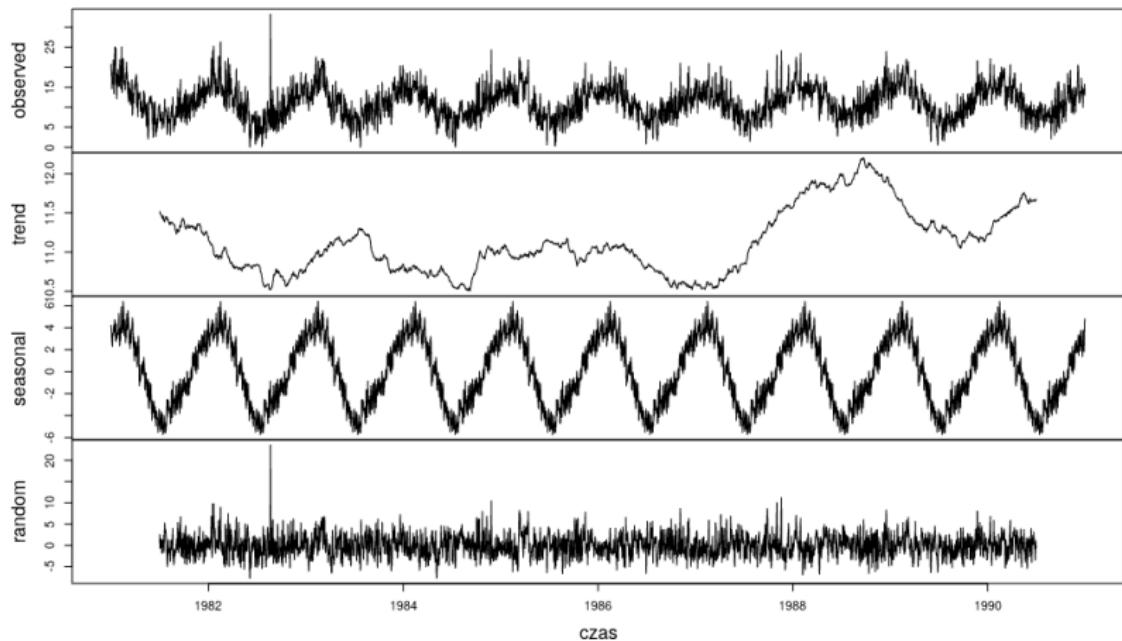
Algorytm K-średnich



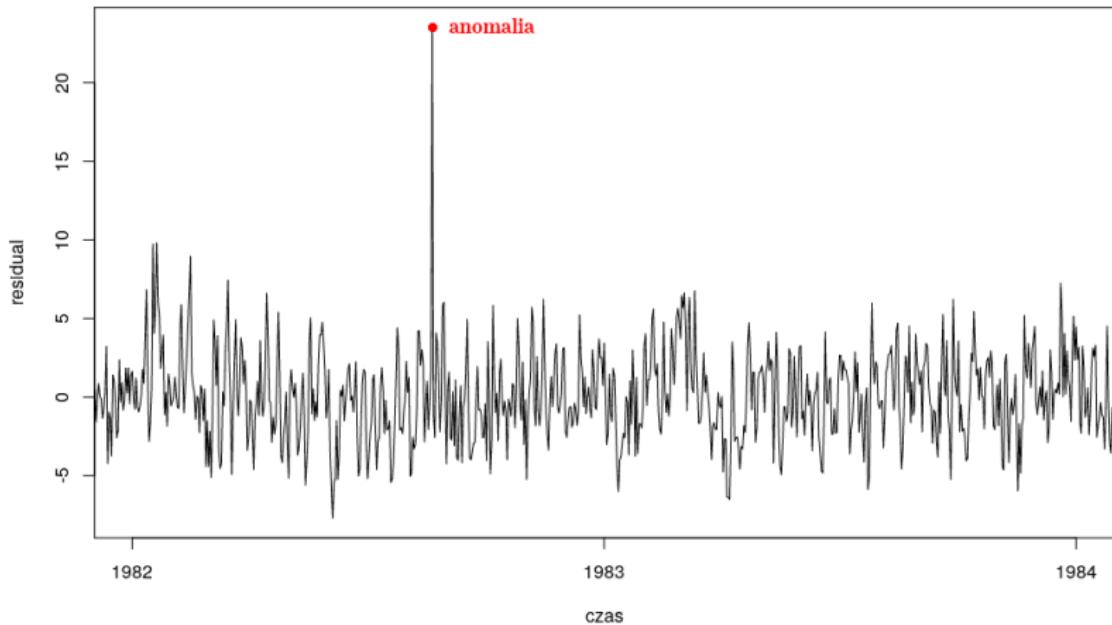


Rysunek: Źródło: <https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e>.

Dekompozycja szeregu czasowego



Dekompozycja szeregu czasowego (c.d.)



1. Agnieszka Landowska, *Materiały wykładowe do przedmiotu Przetwarzanie Danych w Biznesie*, Politechnika Gdańsk, 2020.
2. Aileen Nielsen, *Szeregi czasowe. Praktyczna analiza i predykcja z wykorzystaniem statystyki i uczenia maszynowego*, Helion, 2020.
3. Ane Blázquez-Garcia, Angel Conde, Usue Mori and José Antonio Lozano, *A Review on Outlier/Anomaly Detection in Time Series Data*. ACM Computing Surveys (CSUR) 54 (2021): 1 - 33.
4. Avril Coghlan, *A Little Book of R For Time Series*, 2016,
<https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf>.
5. Robert Nau, *Principles and risks of forecasting*, Fuqua School of Business, Duke University, 2014,
https://people.duke.edu/~rnau/Principles_and_risks_of_forecasting--Robert_Nau.pdf.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Analiza sieciowa

dr inż. Aleksandra Karpus

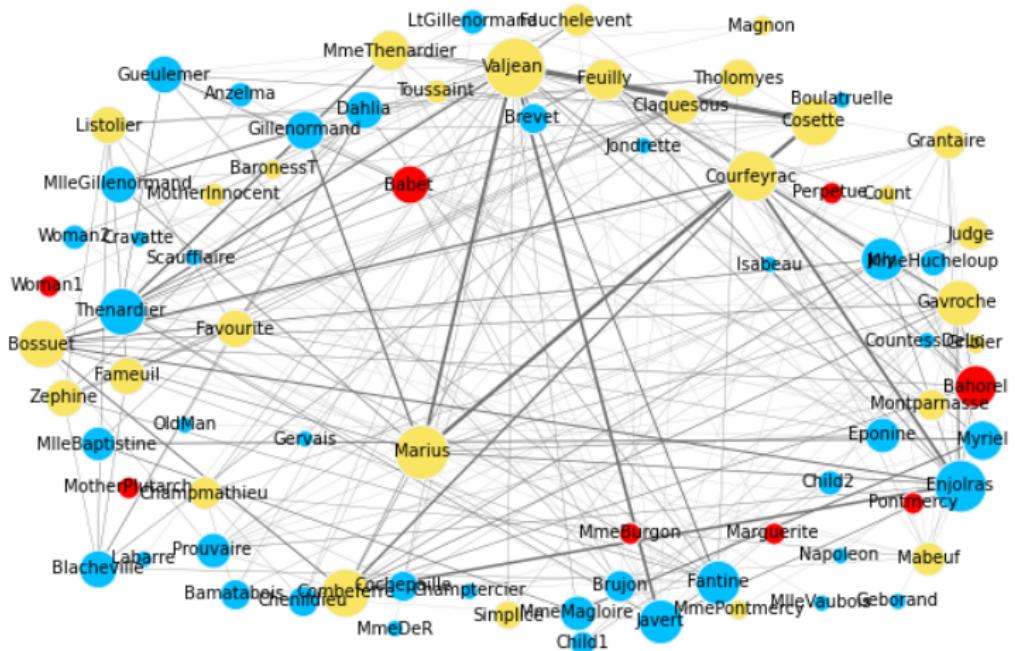
13 grudnia 2023



Czym są sieci?

Sieć

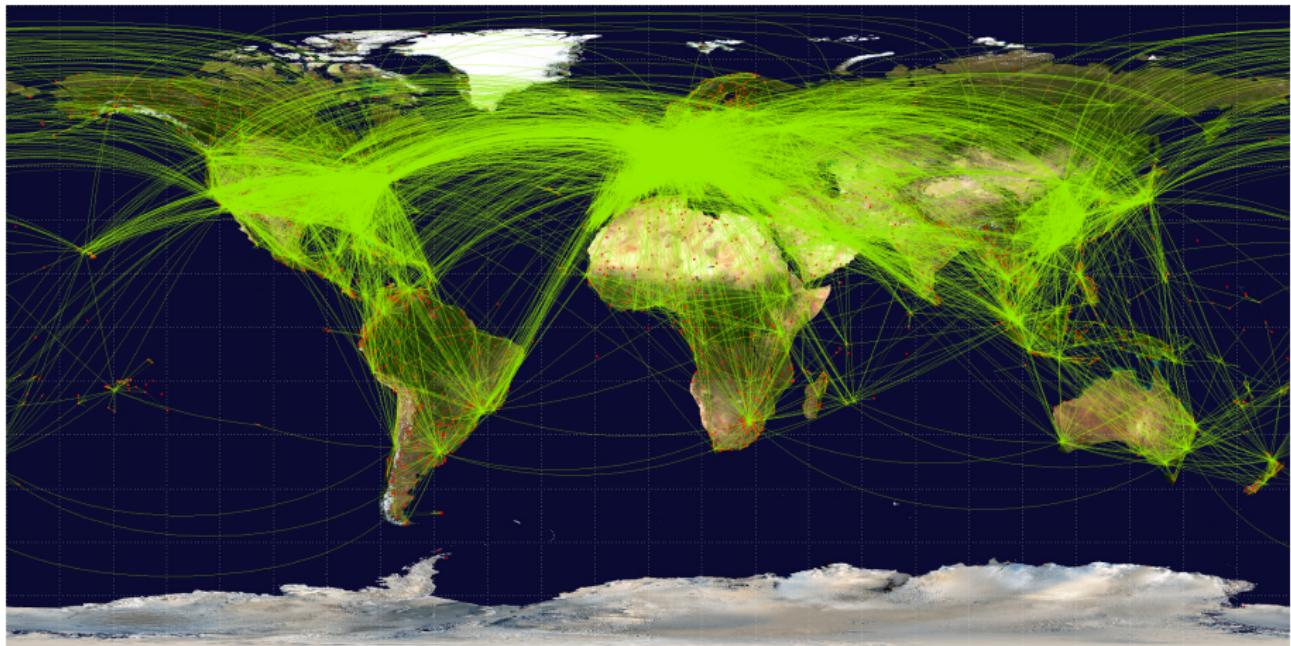
Sieć (lub graf) $G(V,E)$ - zbiór obiektów (reprezentowanych poprzez wierzchołki V) z połączaniami (reprezentowanymi poprzez krawędzie E).



Po co zajmować się analizą sieciową?

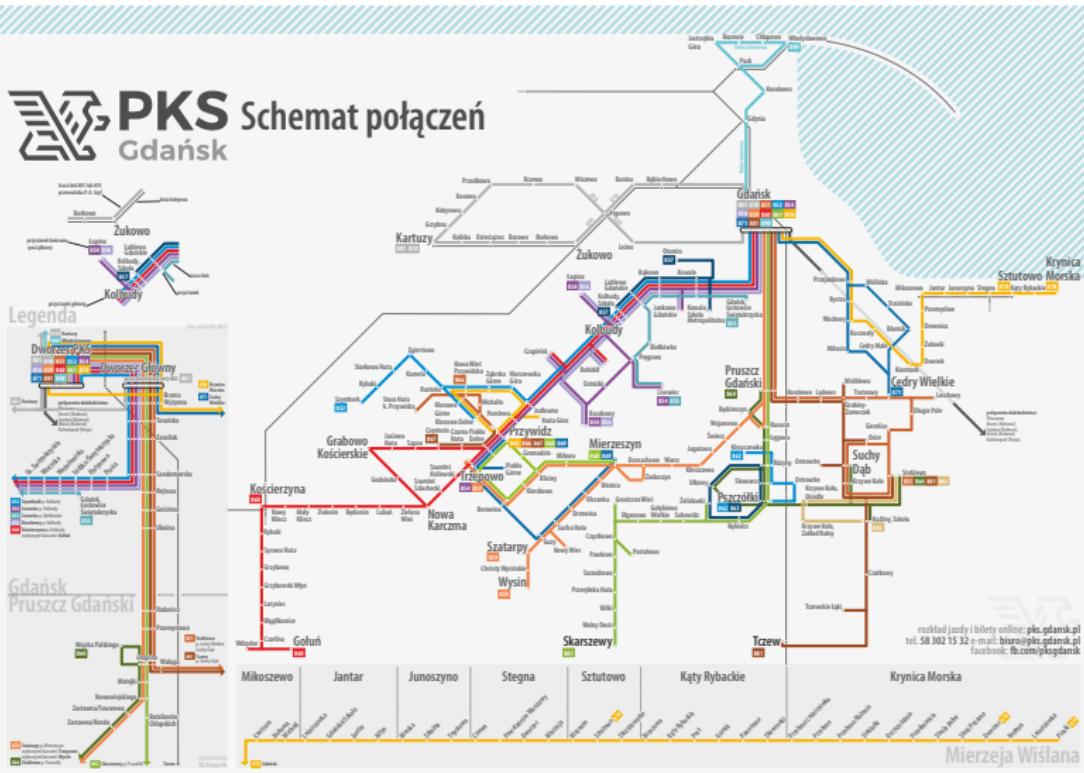
Sieci są wszędzie

Sieć połączeń lotniczych¹



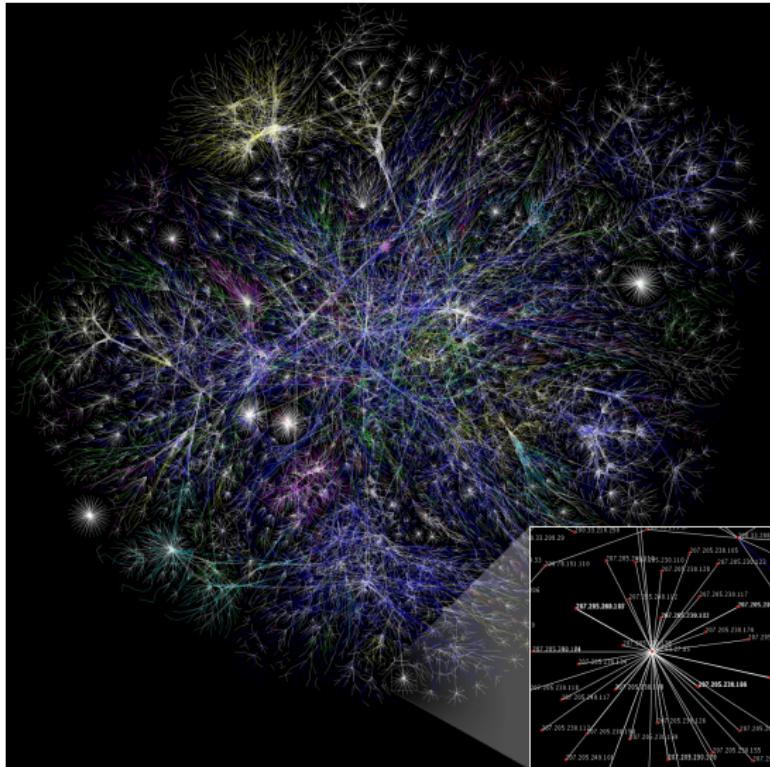
¹<https://upload.wikimedia.org/wikipedia/commons/a/ac/World-airline-routemap-2009.png>

Sieć połączeń autobusowych²



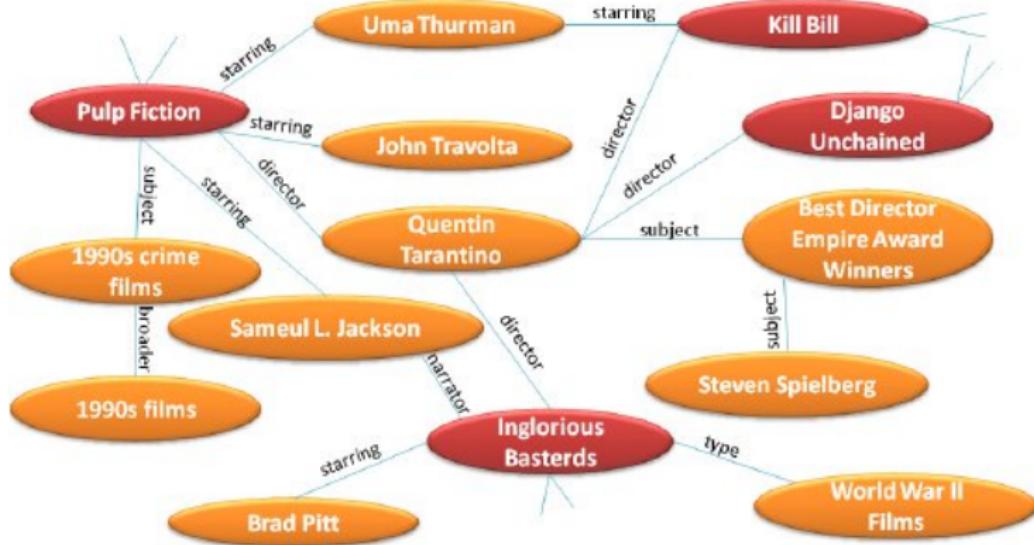
²https://www.pks.gdansk.pl/wp-content/uploads/2023/08/PKSG_schemat-sieci_2023_2.pdf

Sieć Internet³



³https://upload.wikimedia.org/wikipedia/commons/d/d2/Internet_map_1024.jpg

Sieć artykułów w Wikipedii⁴



⁴<https://tiny.pl/9lgxz>

Reakcje zachodzące pomiędzy białkami⁵

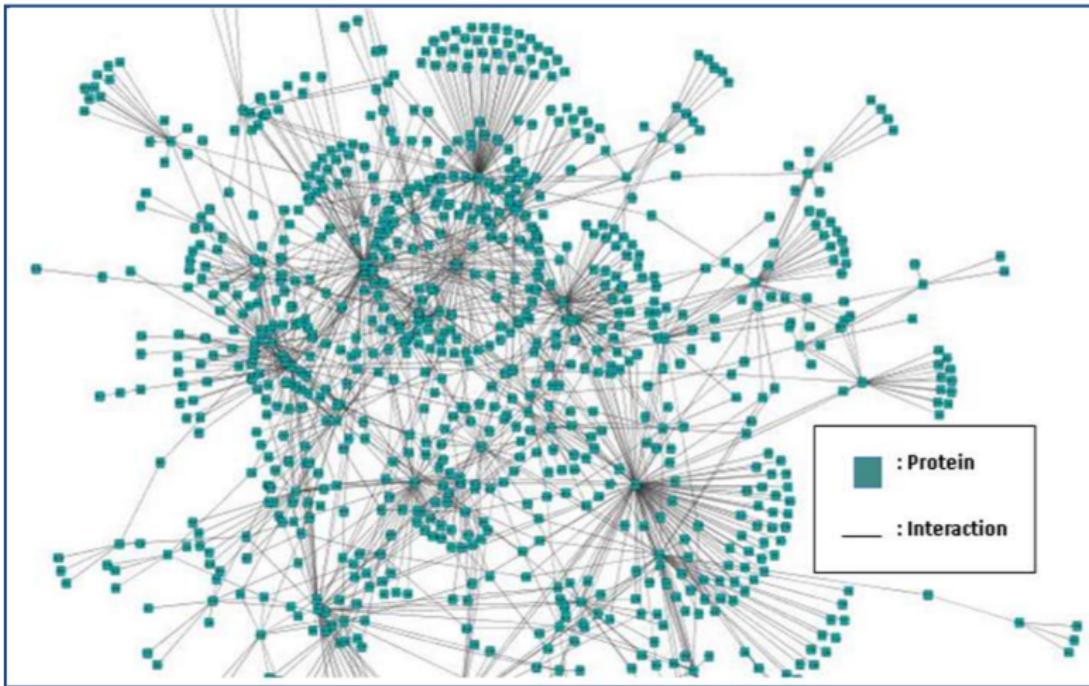
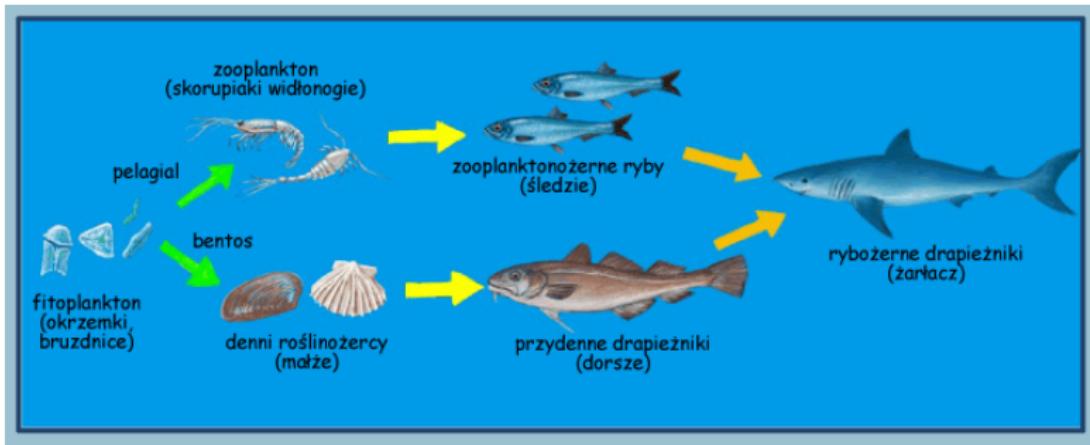


Figure 2: Protein-protein interaction network for Alzheimer's disease

⁵<https://tiny.pl/9lgx3>

Rao, V. Srinivasa et al. "Protein interaction network for Alzheimer's disease using computational approach." Bioinformation 9 (2013): 968 - 972.

Łańcuch pokarmowy zwierząt⁶

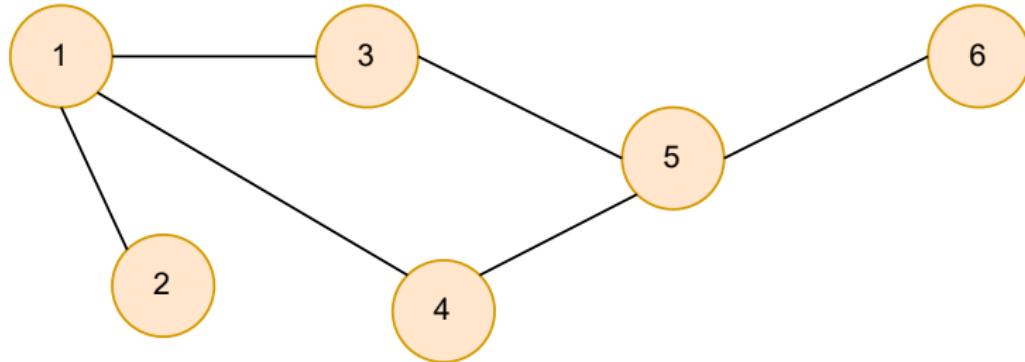


⁶http://muzeum.pgi.gov.pl/lekcje_int/morza/img_srodowisko_biotyczne/5.gif

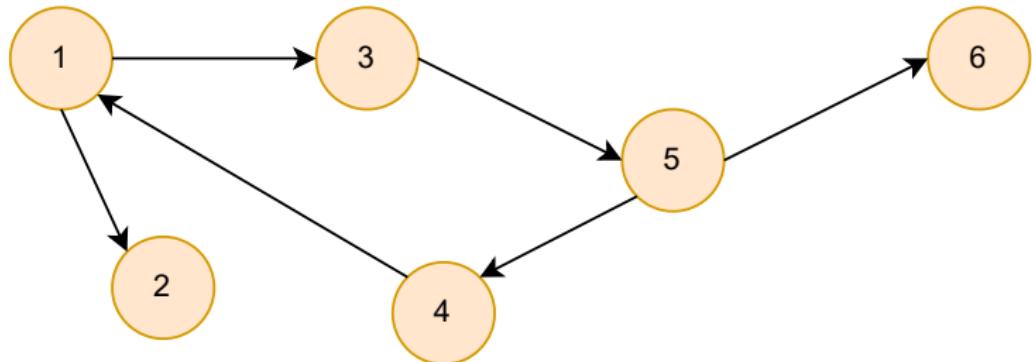
... i wiele innych.

Rodzaje grafów

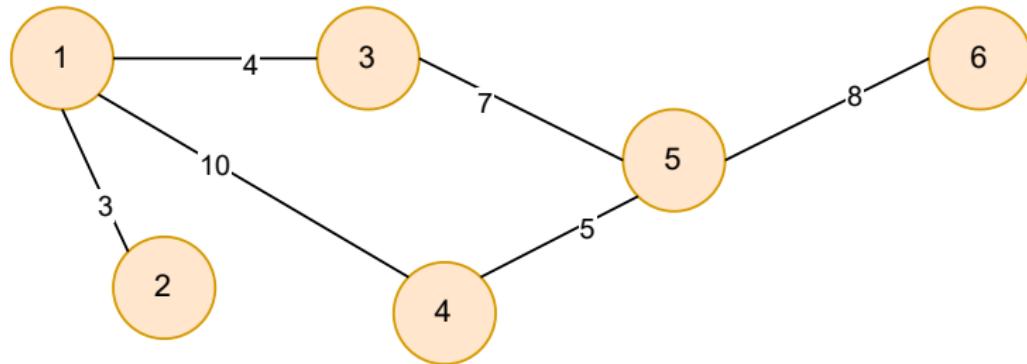
Graf nieskierowany



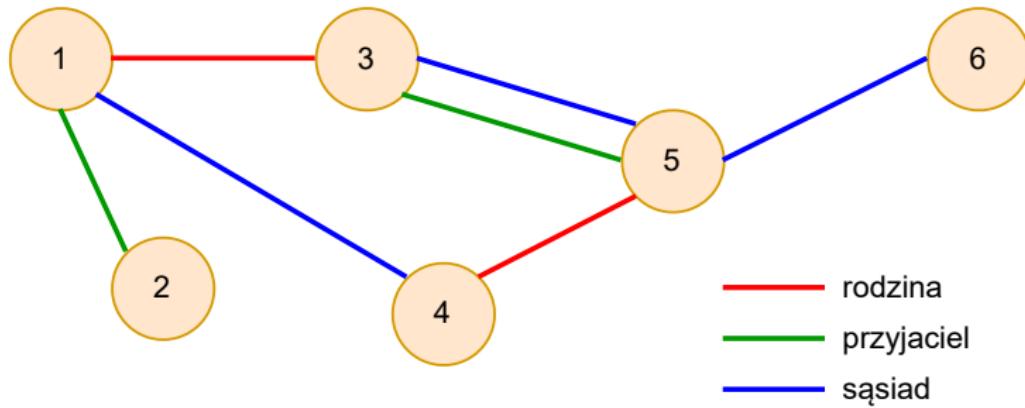
Graf skierowany



Graf ważony



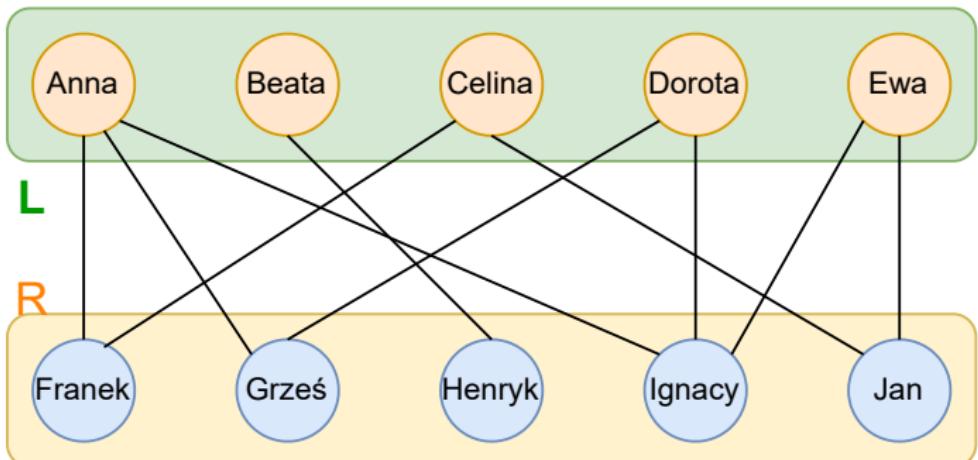
Multigraf



Graf dwudzielny

Graf dwudzielny

to graf, w którym możemy wydzielić dwa zbiory wierzchołków L i R, a każda krawędź w grafie łączy wierzchołek ze zbioru L z wierzchołkiem ze zbioru R.



Własności grafów

Ścieżka

Dowolna droga pomiędzy dwoma wierzchołkami.

Ścieżka

Dowolna droga pomiędzy dwoma wierzchołkami.

Długość ścieżki

Ilość kroków (krawędzi) w ścieżce.

Ścieżka

Dowolna droga pomiędzy dwoma wierzchołkami.

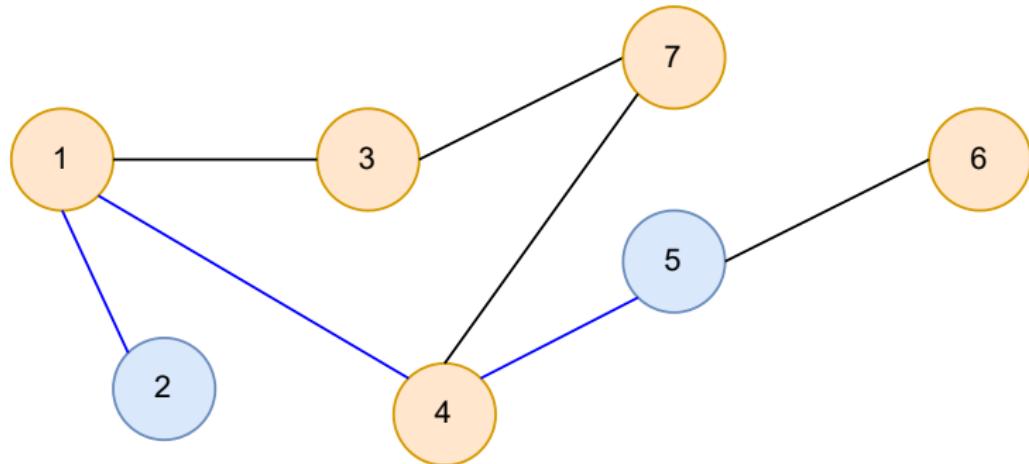
Długość ścieżki

Ilość kroków (krawędzi) w ścieżce.

Odległość

Odlegością dwóch wierzchołków nazywamy długość najkrótszej ścieżki pomiędzy tymi wierzchołkami.

Ścieżki w grafie - przykład



Średnica

Największa odległość pomiędzy dwoma wierzchołkami.

Średnica

Największa odległość pomiędzy dwoma wierzchołkami.

Mimośród wierzchołka

Największa odległość pomiędzy wierzchołkiem i dowolnym innym wierzchołkiem grafu.

Średnica

Największa odległość pomiędzy dwoma wierzchołkami.

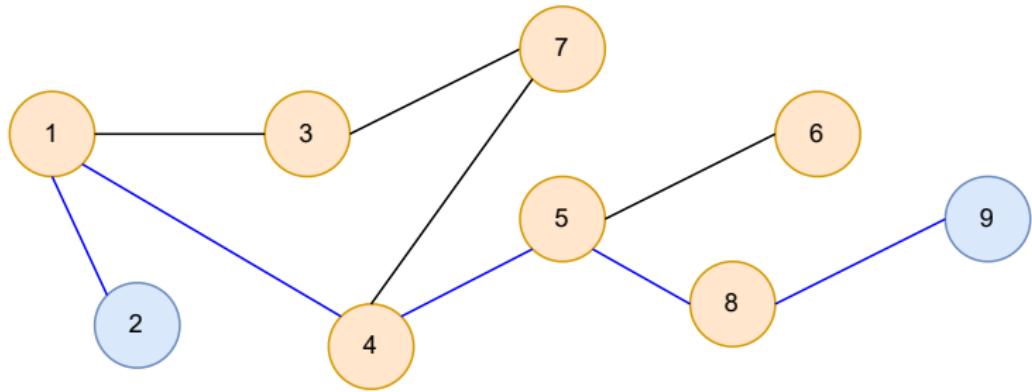
Mimośród wierzchołka

Największa odległość pomiędzy wierzchołkiem i dowolnym innym wierzchołkiem grafu.

Promień

Najmniejszy mimośród w grafie.

Średnica grafu - przykład



Obrzeże grafu

Zbiór wierzchołków, których mimośród jest równy średnicy grafu.

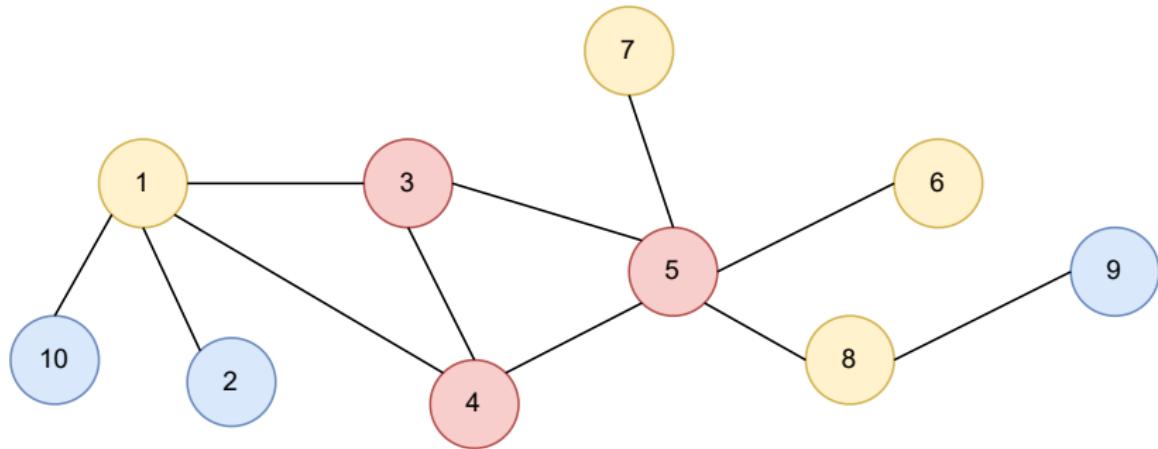
Obrzeże grafu

Zbiór wierzchołków, których mimośród jest równy średnicy grafu.

Centrum grafu

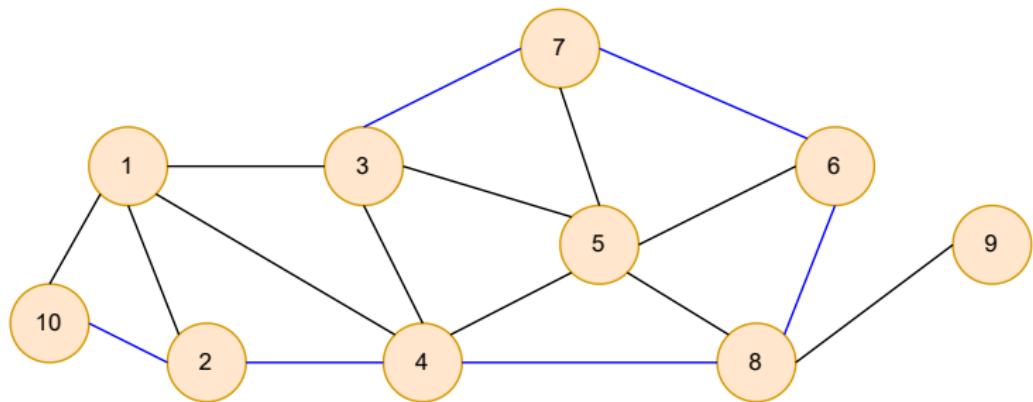
Zbiór wierzchołków, których mimośród jest równy promieniowi grafu.

Centrum i obrzeże - przykład



Dopełnienie triadyczne (ang. *Triadic closure*)

to tendencja, że wierzchołki, które mają wspólne połączenie, również będą połączone w przyszłości.



- "Przyjaciel mojego przyjaciela jest moim przyjacielem"
- "Wróg mojego wroga jest moim przyjacielem"

Lokalny współczynnik gronowania (ang. *Local Clustering Coefficient*)

Stosunek par przyjaciół węzła, którzy są swoimi przyjaciółmi do wszystkich par przyjaciół tego węzła.

$$\text{LCC} = \frac{\#\text{par przyjaciół węzła, którzy są swoimi przyjaciółmi}}{\#\text{par przyjaciół węzła}}$$

Globalny współczynnik gronowania (ang. *Global Clustering Coefficient*)

Uśredniony lokalny współczynnik gronowania dla wszystkich wierzchołków grafu.

Graf spójny

Taki graf nieskierowany, w którym istnieje ścieżka pomiędzy każdą parą wierzchołków.

Graf spójny

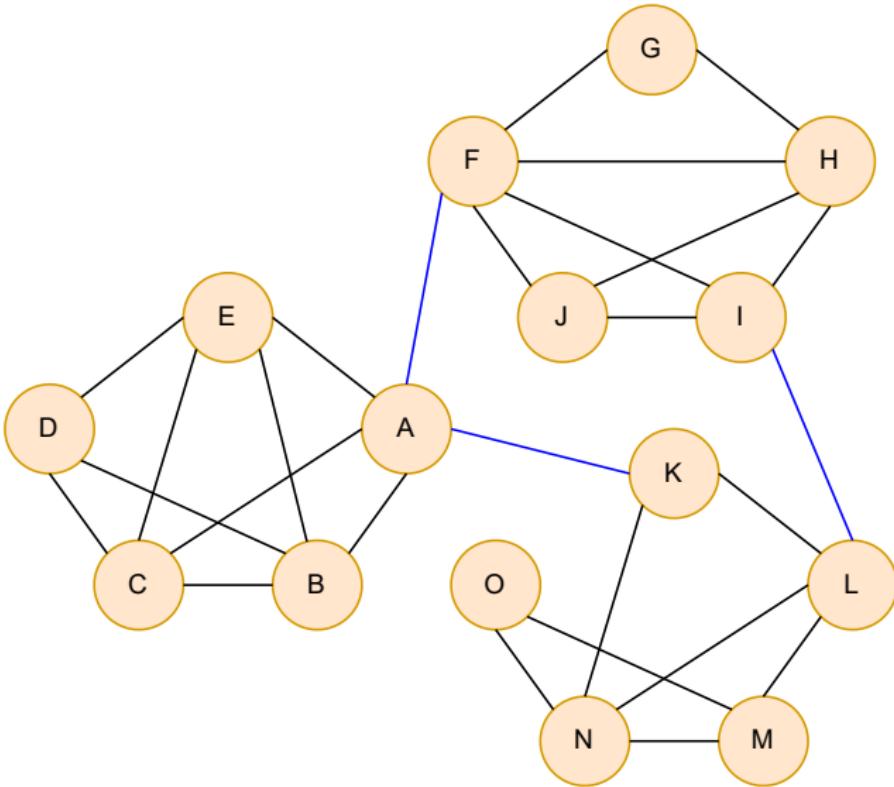
Taki graf nieskierowany, w którym istnieje ścieżka pomiędzy każdą parą wierzchołków.

Spójna składowa grafu

Podzbiór wierzchołków grafu nieskierowanego, dla których:

- istnieje ścieżka pomiędzy każdą parą wierzchołków w danym podzbiorze,
- nie istnieje ścieżka łącząca dowolny wierzchołek tego podzbioru z innym wierzchołkiem grafu.

Spójność grafu - przykład



A co w przypadku grafów skierowanych?

Graf silnie spójny

Taki graf skierowany, w którym istnieje skierowana ścieżka pomiędzy wierzchołkami u i v oraz v i u .

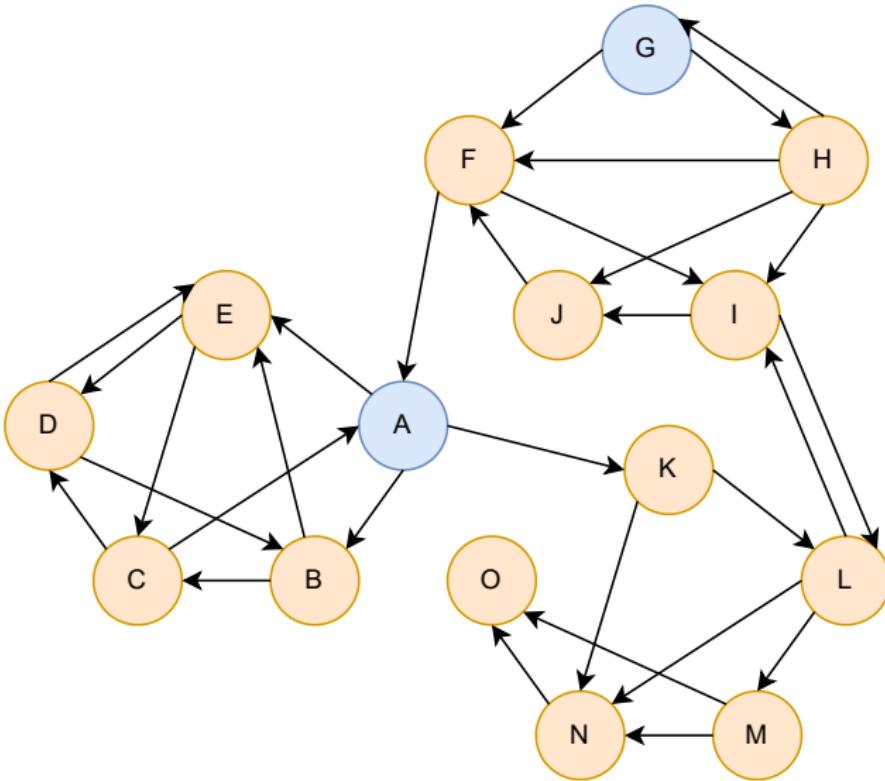
Graf silnie spójny

Taki graf skierowany, w którym istnieje skierowana ścieżka pomiędzy wierzchołkami u i v oraz v i u .

Graf słabo spójny

Taki graf skierowany, w którym po zmianie krawędzi skierowanych na nieskierowane, graf jest spójny.

Spójność grafu skierowanego - przykład



Silna spójna składowa grafu

Podzbiór wierzchołków grafu skierowanego, dla których:

- istnieje skierowana ścieżka pomiędzy każdą parą wierzchołków w danym podzbiorze,
- nie istnieje skierowana ścieżka łącząca dowolny wierzchołek tego podzbioru z innym wierzchołkiem grafu.

Silna spójna składowa grafu

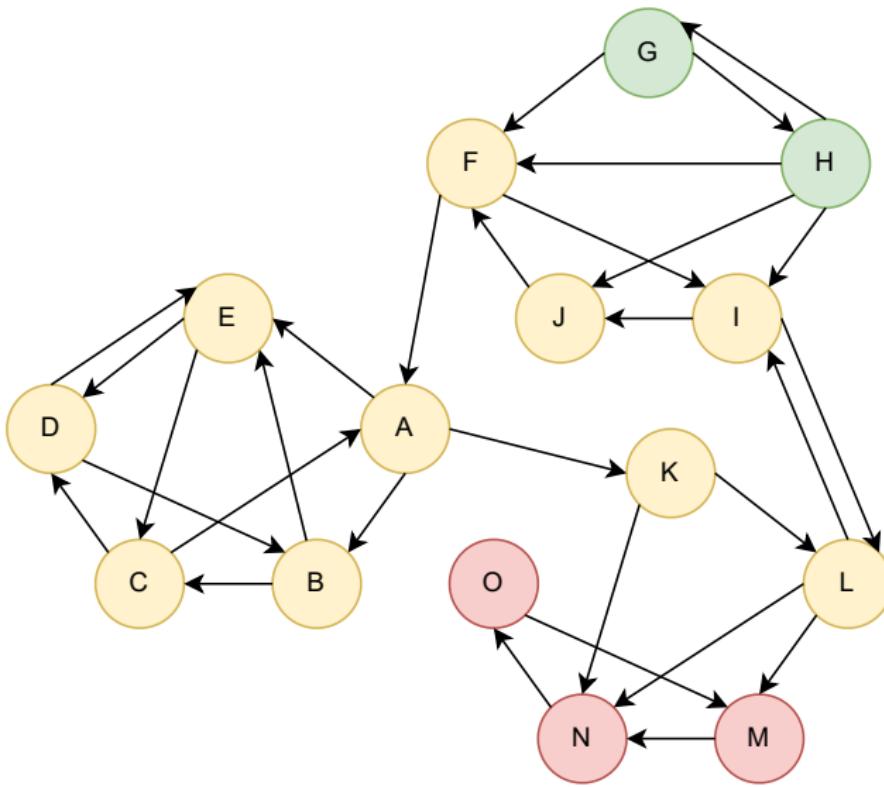
Podzbiór wierzchołków grafu skierowanego, dla których:

- istnieje skierowana ścieżka pomiędzy każdą parą wierzchołków w danym podzbiorze,
- nie istnieje skierowana ścieżka łącząca dowolny wierzchołek tego podzbioru z innym wierzchołkiem grafu.

Słaba spójna składowa grafu

Spójna składowa grafu po zastąpieniu wszystkich krawędzi skierowanych krawędziami nieskierowanymi.

Silna spójna składowa grafu skierowanego - przykład



Odporność/niezawodność sieci

Co to może znaczyć, że sieć jest odporna/niezawodna?

Odporność sieci (ang. *Network robustness*)

zdolność sieci do zachowania swoich ogólnych właściwości strukturalnych w przypadku awarii lub ataków.

Odporność sieci (ang. *Network robustness*)

zdolność sieci do zachowania swoich ogólnych właściwości strukturalnych w przypadku awarii lub ataków.

Typy ataków:

- Usunięcie wierzchołków.
- Usunięcie krawędzi.

Strukturalne właściwości:

spójność grafu/sieci.

Rozspojnianie grafu

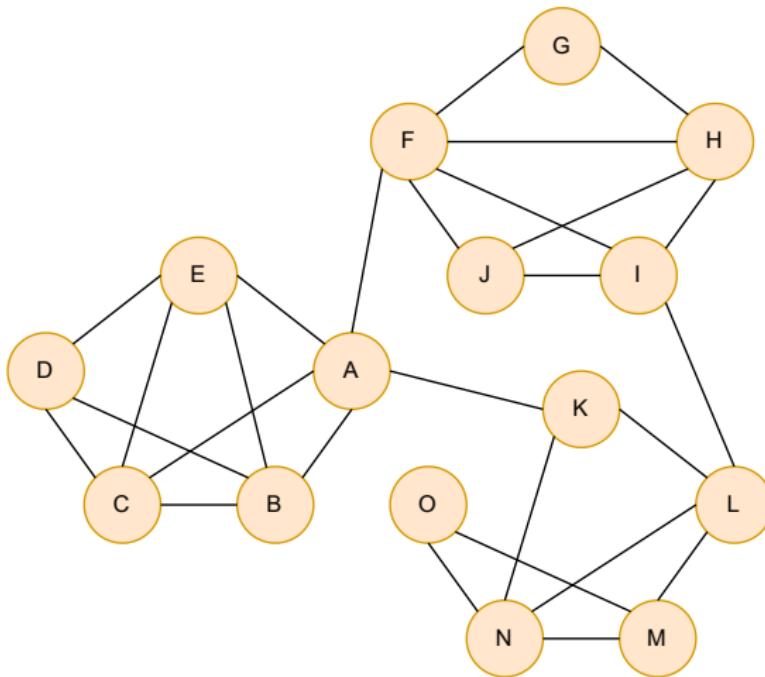
Graf/sieć możemy rozspojnić poprzez usunięcie:

- wierzchołków.
- krawędzi.

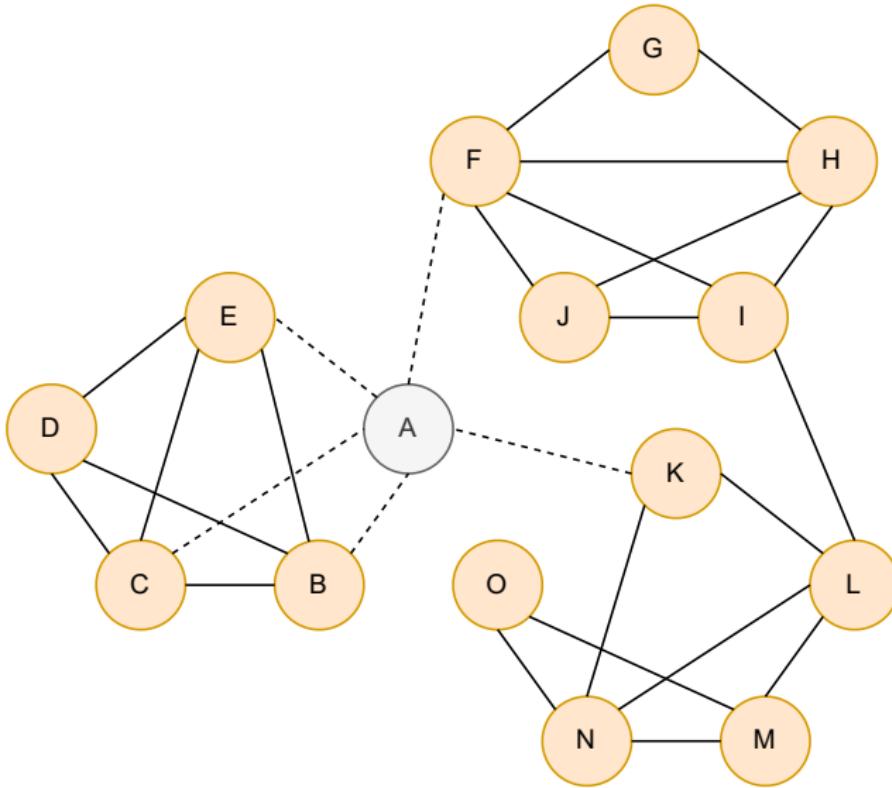
Rozspójnenie grafu

- usuwanie wierzchołków

Jaka jest najmniejsza liczba wierzchołków, które należy usunąć, aby graf stał się niespójny?



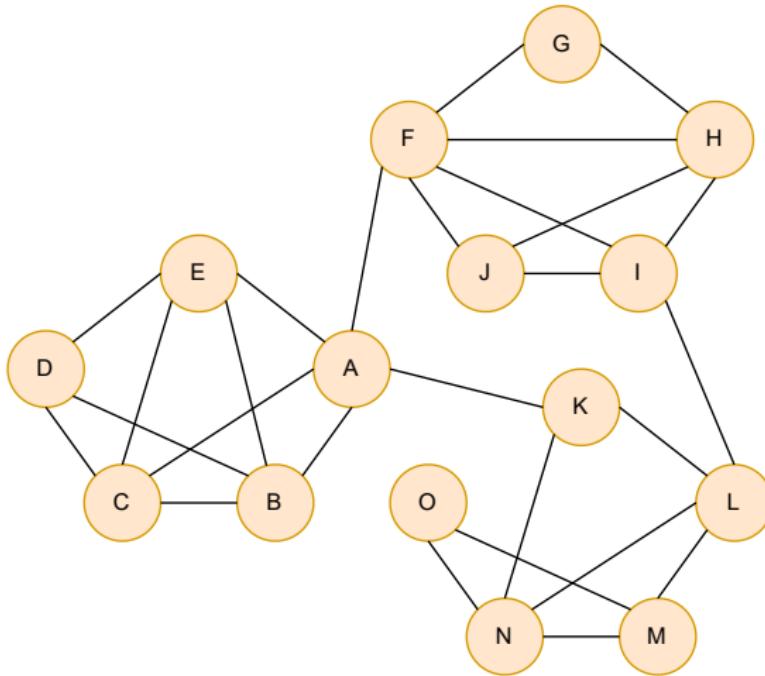
Rozspójnianie grafu - przykład



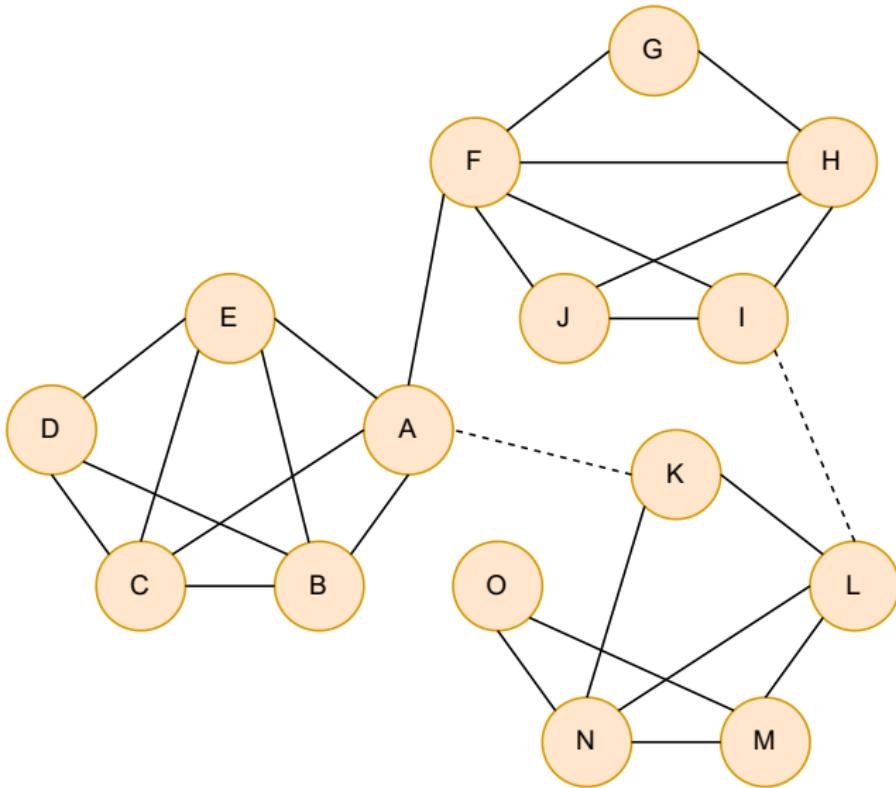
Rozspójnianie grafu

- usuwanie krawędzi

Jaka jest najmniejsza liczba krawędzi, które należy usunąć, aby graf stał się niespójny?



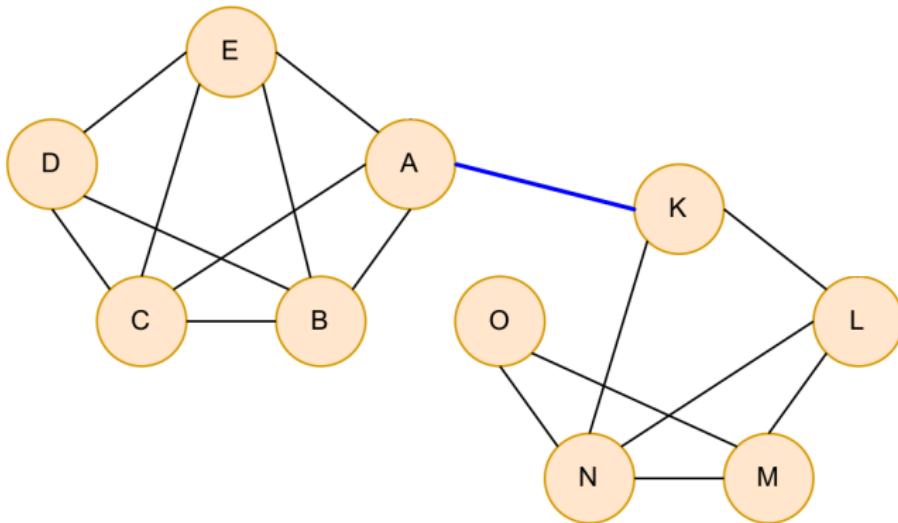
Rozspojnianie grafu - przykład 2



Mostem określamy taką krawędź AB, po której usunięciu wierzchołki A i B leżą w różnych spójnych składowych grafu.

Most (ang. *bridge*)

Mostem określamy taką krawędź AB, po której usunięciu wierzchołki A i B leżą w różnych spójnych składowych grafu.



Co to może znaczyć, że sieć jest odporna/niezawodna?

Odporność sieci (2)

Oporne sieci mają duże minimalne cięcia węzłów i krawędzi, tzn. należy usunąć dużą liczbę krawędzi lub wierzchołków, aby rozspojnić sieć.

W jaki sposób możemy określić istotność wierzchołka w sieci?

Możemy ustalić istotność wierzchołka, np. poprzez:

- stopień wierzchołka, np. liczba przyjaciół,
- średnia bliskość do innych węzłów,
- procent najkrótszych ścieżek, które przechodzą przez węzeł.

Miary centralności identyfikują najważniejsze węzły w sieci:

- Wpływowe węzły w sieci społecznościowej.
- Węzły, które rozpowszechniają informacje do wielu węzłów lub zapobiegają epidemiom.
- Węzły w sieci transportowej.
- Ważne strony w sieci.
- Węzły zapobiegające rozpadowi sieci.

- Centralność stopnia
- Centralność bliskości
- Centralność pośredniczości
- Centralność ładowania
- Page Rank
- centralność Katza

Centralność stopnia

- Założenie: istotne wierzchołki mają dużo połączeń.
- Najprostsza miara centralności - wykorzystuje liczbę sąsiadów.

$$C_{deg}(v) = \frac{d_v}{|N|-1},$$

gdzie d_v to stopień wierzchołka v , zaś N to zbiór wszystkich wierzchołków w grafie.

- Dla grafów nieskierowanych wykorzystujemy stopień wierzchołka.
- Dla grafów skierowanych wykorzystujemy stopień wejściowy lub wyjściowy.

- Założenie: istotne wierzchołki leżą blisko innych wierzchołków.

$$C_{close}(v) = \frac{|N| - 1}{\sum_{u \in N-v} d(v, u)},$$

gdzie N - zbiór wierzchołków, $d(v, u)$ - długość najkrótszej ścieżki z v do u .

- Założenie: istotne wierzchołki łączą inne wierzchołki.

$$C_{btw}(v) = \sum_{s,t \in N} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}},$$

gdzie $\sigma_{s,t}$ - liczba najkrótszych ścieżek pomiędzy s i t , $\sigma_{s,t}(v)$ - liczba najkrótszych ścieżek pomiędzy s i t , które przechodzą przez v .

1. Daniel Romero, *Applied Social Network Analysis in Python*, University of Michigan, Coursera, 2020.
2. John Scott, *Social Network Analysis*, 4th edition, Sage Publications, 2017.

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Odkrywanie Wiedzy i Systemy Rekomendacyjne

Analiza sieciowa

PageRank i fenomen "Małego Świata"

dr inż. Aleksandra Karpus

20 grudnia 2023

Agenda

1. Miary centralności
2. Algorytm PageRank
3. Fenomen "Małego Świata"

Co to może znaczyć, że sieć jest odporna/niezawodna?

Oporne sieci mają duże minimalne cięcia węzłów i krawędzi, tzn. należy usunąć dużą liczbę krawędzi lub wierzchołków, aby rozspojnić sieć.

W jaki sposób możemy określić istotność wierzchołka w sieci?

Możemy ustalić istotność wierzchołka, np. poprzez:

- stopień wierzchołka, np. liczba przyjaciół,
- średnia bliskość do innych węzłów,
- procent najkrótszych ścieżek, które przechodzą przez węzeł.

Miary centralności identyfikują najważniejsze węzły w sieci:

- Wpływowe węzły w sieci społecznościowej.
- Węzły, które rozpowszechniają informacje do wielu węzłów lub zapobiegają epidemiom.
- Węzły w sieci transportowej.
- Ważne strony w sieci.
- Węzły zapobiegające rozpadowi sieci.

- Centralność stopnia
- Centralność bliskości
- Centralność pośredniczości
- Centralność ładowania
- Page Rank
- centralność Katza

- Założenie: istotne wierzchołki mają dużo połączeń.
- Najprostsza miara centralności - wykorzystuje liczbę sąsiadów.

$$C_{deg}(v) = \frac{d_v}{|N|-1},$$

gdzie d_v to stopień wierzchołka v , zaś N to zbiór wszystkich wierzchołków w grafie.

- Dla grafów nieskierowanych wykorzystujemy stopień wierzchołka.
- Dla grafów skierowanych wykorzystujemy stopień wejściowy lub wyjściowy.

- Założenie: istotne wierzchołki leżą blisko innych wierzchołków.

$$C_{close}(v) = \frac{|N| - 1}{\sum_{u \in N-v} d(v, u)},$$

gdzie N - zbiór wierzchołków, $d(v, u)$ - długość najkrótszej ścieżki z v do u .

- Założenie: istotne wierzchołki łączą inne wierzchołki.

$$C_{btw}(v) = \sum_{s,t \in N} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}},$$

gdzie $\sigma_{s,t}$ - liczba najkrótszych ścieżek pomiędzy s i t , $\sigma_{s,t}(v)$ - liczba najkrótszych ścieżek pomiędzy s i t , które przechodzą przez v .

- Algorytm wymyślony przez twórców wyszukiwarki Google, Lawrence'a Page'a i Sergey'a Brina, w 1989 roku.
- Algorytm ten przysłużył się do ogromnej popularności Google.
- PageRank przypisuje każdej stronie internetowej pewną ocenę będącą ważnością danej strony.
- Ważność jest definiowana rekurencyjnie poprzez ważność stron, które linkują do tej strony.
- Oryginalnie wymyślony do stron internetowych, jednak może być wykorzystywany dla każdego rodzaju sieci, zwłaszcza skierowanych.

Algorytm PageRank

– wersja podstawowa

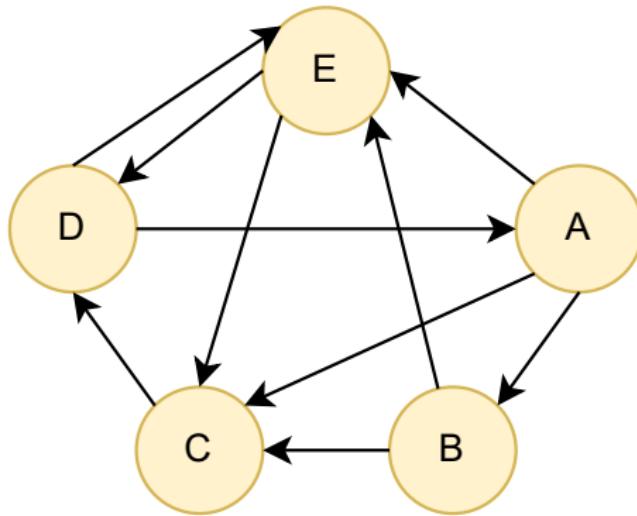
n - liczba wierzchołków w sieci

k - liczba kroków

1. Przypisz każdemu wierzchołkowi ważność równą $\frac{1}{n}$.
2. Wykonaj k razy:
 1. Wyznacz przekazywaną przez wierzchołek x ważność dla każdego wierzchołka y , do którego linkuje, jako ważność wierzchołka x podzieloną przez liczbę krawędzi wychodzących z tego wierzchołka.
 2. Wyznacz nową wartość ważności dla wierzchołka y jako sumę ważności przekazywanych przez wszystkie wierzchołki, które do niego linkują.
3. Zwróć sieć z przypisaną ważnością wierzchołków.

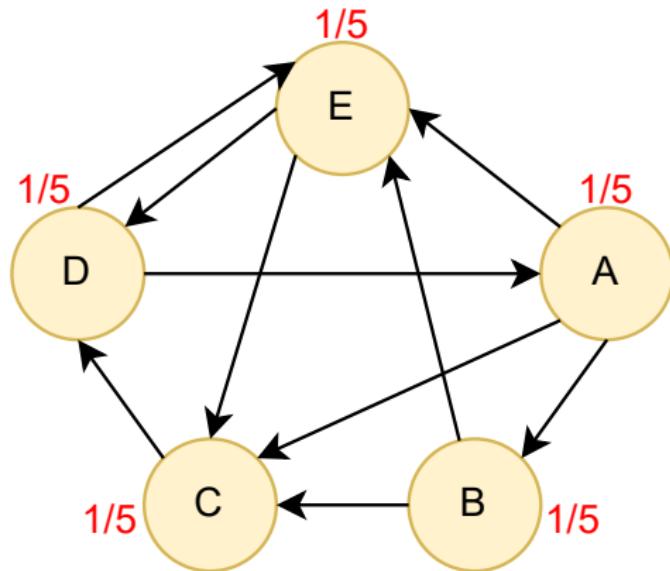
Algorytm PageRank – przykład

Wyznaczmy ważność wierzchołków po dwóch iteracjach algorytmu PageRank dla poniższej sieci.



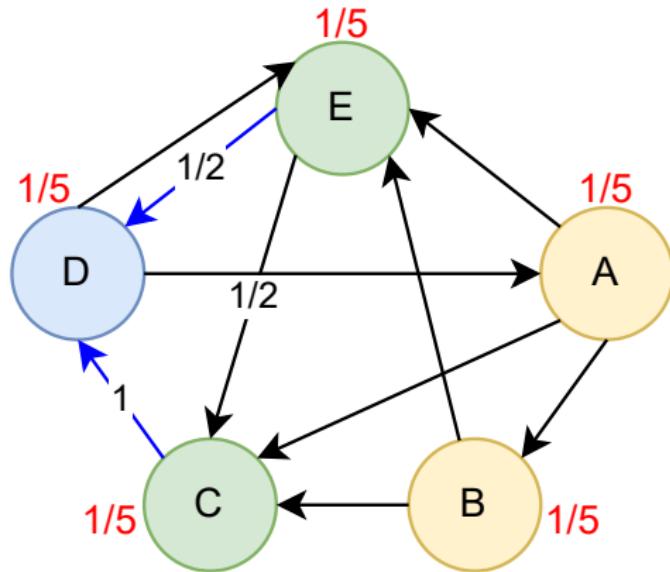
Algorytm PageRank – przykład

Krok 1 - Przypisanie inicjalnych wartości dla ważności.



Algorytm PageRank – przykład

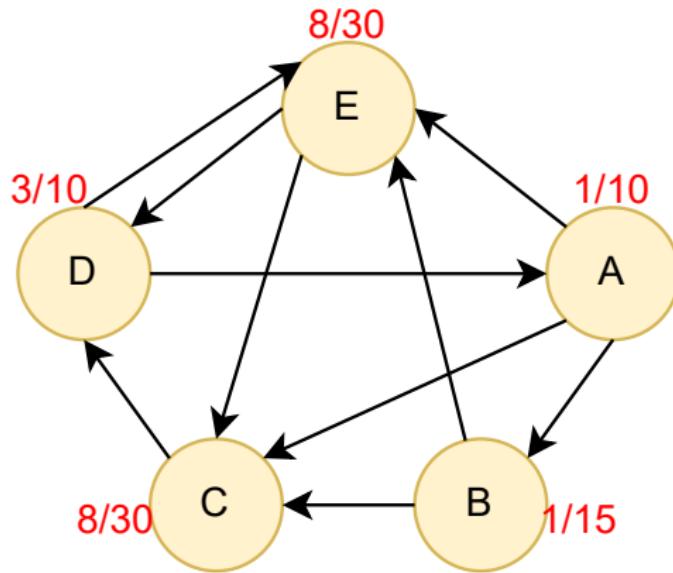
Iteracja 1 - Wyznaczenie ważności dla wierzchołka D.



$$D: \frac{1}{2} * \frac{1}{5} + 1 * \frac{1}{5} = \frac{3}{10}$$

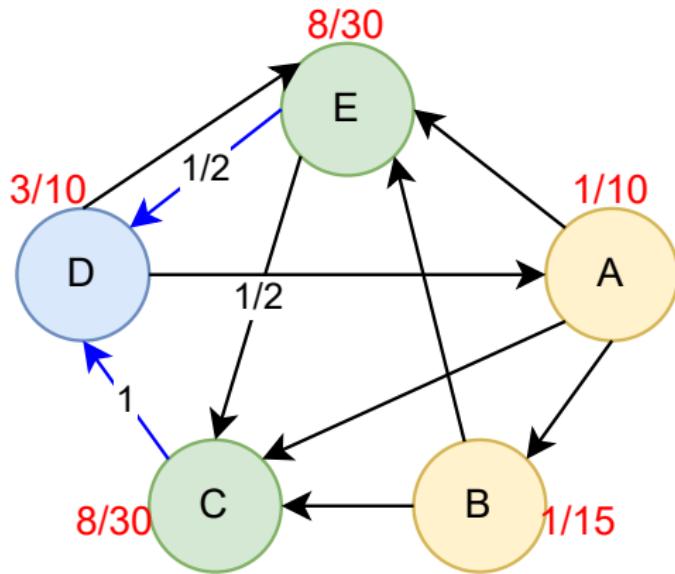
Algorytm PageRank – przykład

Iteracja 1 - Wyznaczone ważności dla wszystkich wierzchołków.



Algorytm PageRank – przykład

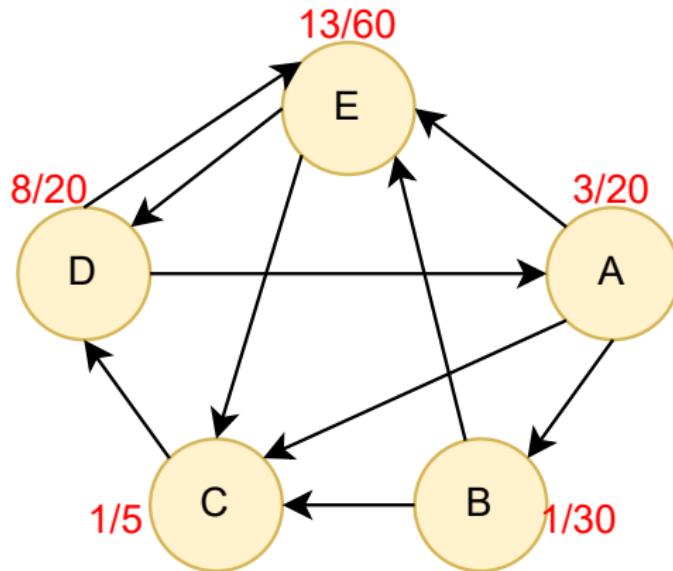
Iteracja 2 - Wyznaczenie ważności dla wierzchołka D.



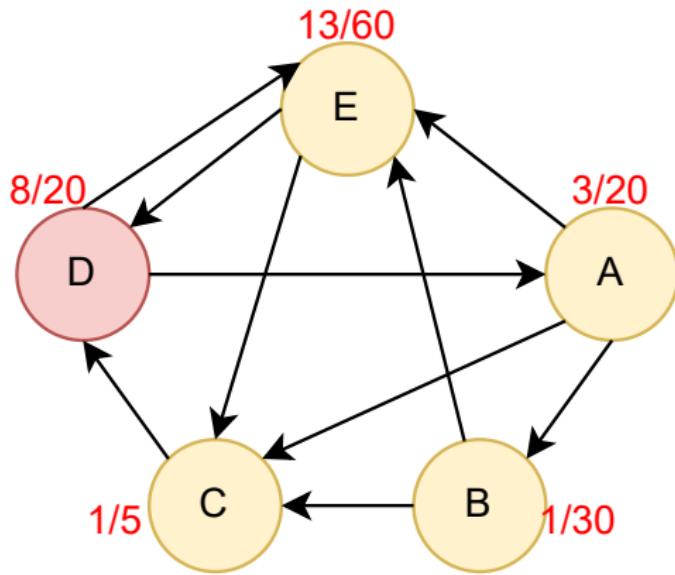
$$D: \frac{1}{2} * \frac{8}{30} + 1 * \frac{8}{30} = \frac{8}{20}$$

Algorytm PageRank – przykład

Iteracja 2 - Wyznaczone ważności dla wszystkich wierzchołków.



Algorytm PageRank – przykład



A co będzie dla $k = 3, 4, 5, \dots$?

A co będzie dla $k = 3, 4, 5, \dots$?

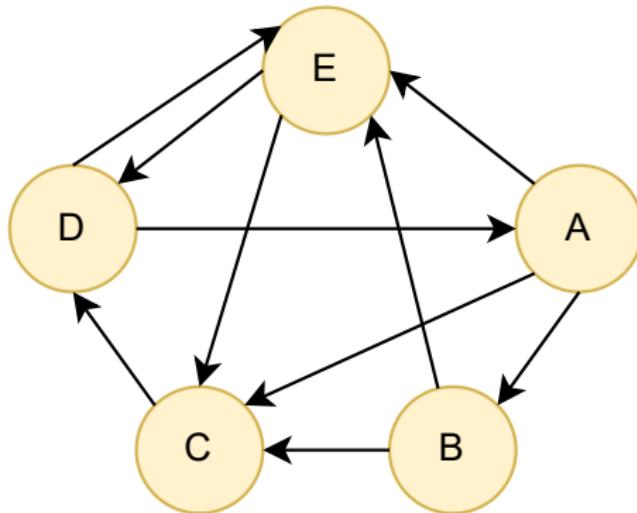
Dla większości sieci ważność jest zbieżna do pewnej wartości przy $k \rightarrow \infty$.

Algorytm PageRank – interpretacja

spacer losowy

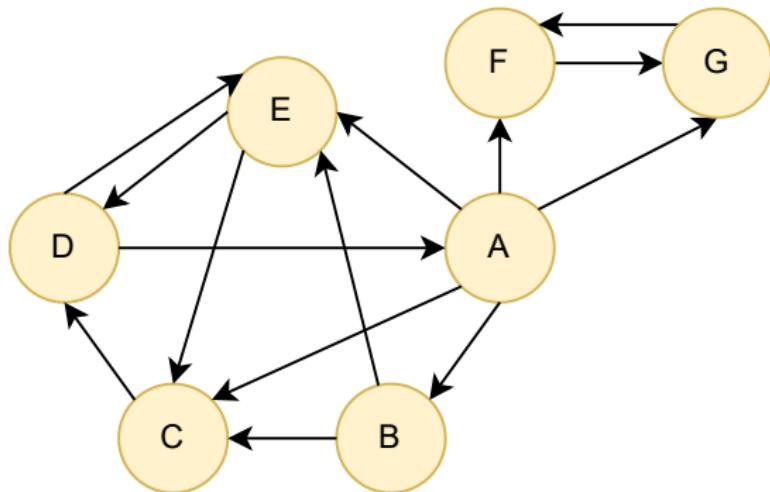
Algorytm PageRank – interpretacja

spacer losowy



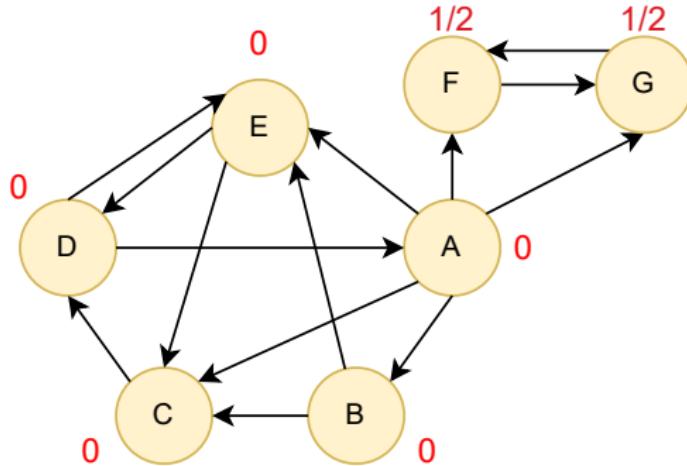
Algorytm PageRank – przykład 2

Do jakich wartości będzie zbiegać ważność po k iteracjach dla poniższej sieci?



Algorytm PageRank – przykład 2 (c.d.)

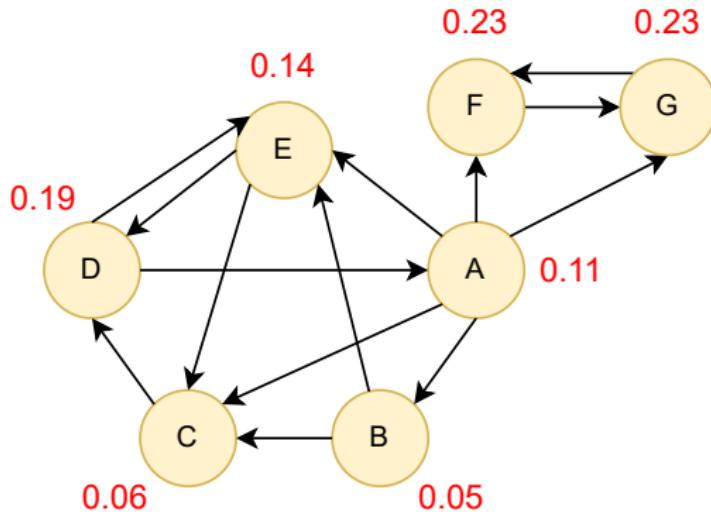
Ważność wierzchołków sieci po $k \rightarrow \infty$ iteracjach algorytmu PageRank:



- Rozszerzony o współczynnik tłumienia α .
- Z prawdopodobieństwem α wykonujemy kroki zgodnie z algorytmem PageRank.
- Z prawdopodobieństwem równym $1 - \alpha$ skaczemy do losowo wybranego wierzchołka.
- Wartości dla parametru α są zwykle wybierane z przedziału $[0.8, 0.9]$.
- Minimalizujemy możliwość "utknięcia" w jednym obszarze sieci.

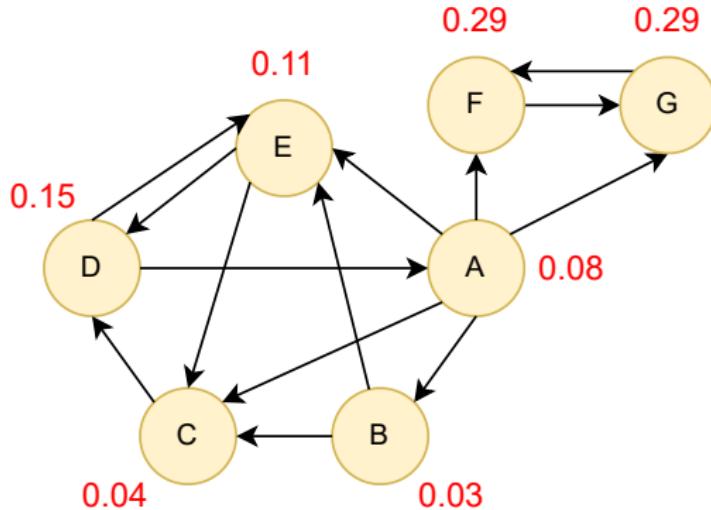
Skalowany PageRank – przykład 3

Ważność wierzchołków sieci po $k \rightarrow \infty$ iteracjach algorytmu PageRank ze współczynnikiem tłumienia $\alpha = 0.8$:



Skalowany PageRank – przykład 4

Ważność wierzchołków sieci po $k \rightarrow \infty$ iteracjach algorytmu PageRank ze współczynnikiem tłumienia $\alpha = 0.9$:



Fenomen Małego Świata

O Małym Świecie mówimy wtedy, jeśli w sieci istnieją krótkie ścieżki pomiędzy prawie każdą parą wierzchołków.

- W latach 60-tych ubiegłego wieku.
- 296 losowo wybranych osób początkowych.
- Zadanie: dostarczyć list do maklera giełdowego mieszkającego na przedmieściach Bostonu.
- Warunek: list mogą wysłać tylko osobie, którą znają, ale mogą przekazać dalej list z instrukcjami.

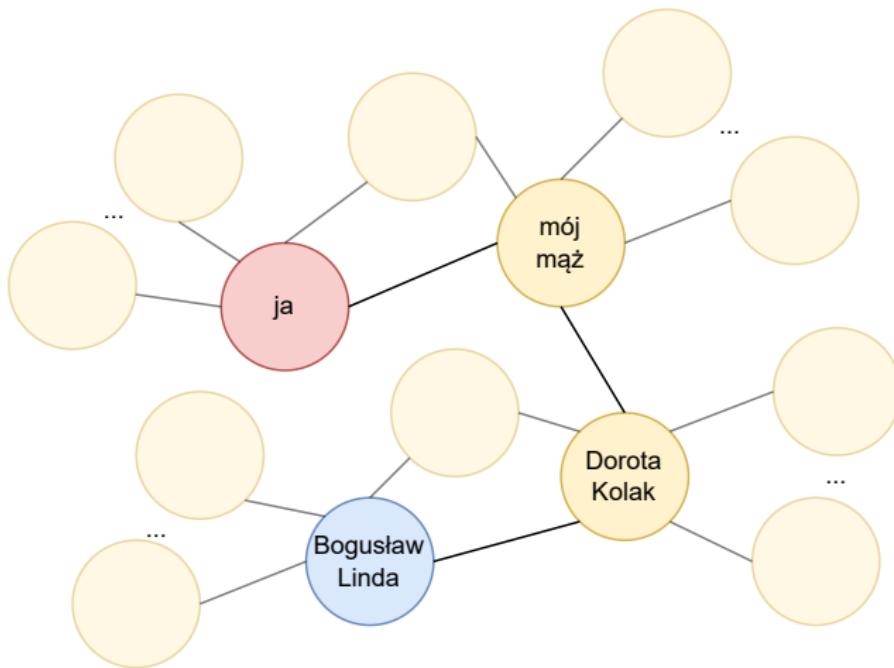
- W latach 60-tych ubiegłego wieku.
- 296 losowo wybranych osób początkowych.
- Zadanie: dostarczyć list do maklera giełdowego mieszkającego na przedmieściach Bostonu.
- Warunek: list mogą wysłać tylko osobie, którą znają, ale mogą przekazać dalej list z instrukcjami.

Wyniki:

- 64 listy dotarły do docelowego adresata.
- Mediana długości ścieżki wynosiła 6.

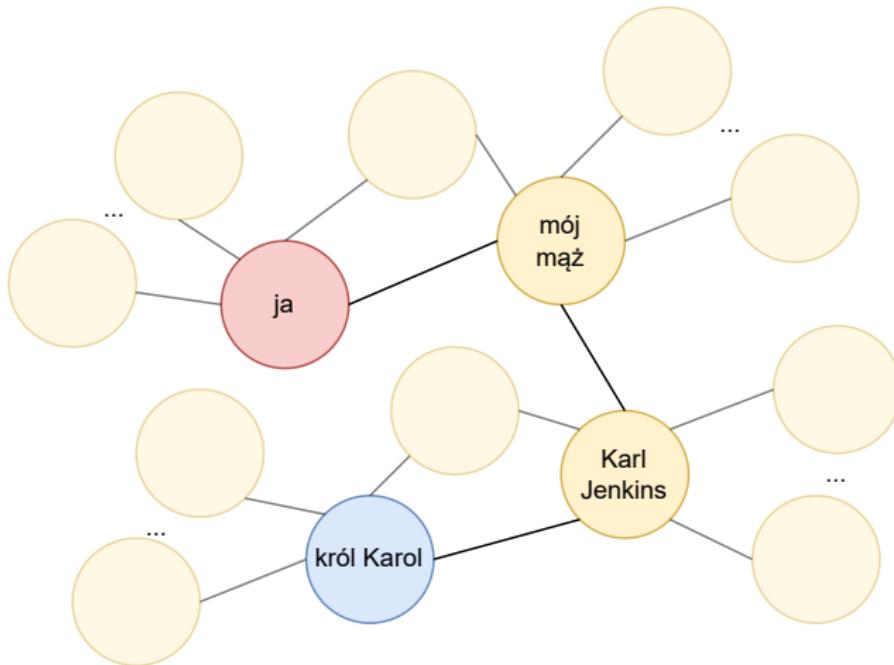
Mały Świat – przykład

Moja znajomość z Bogusławem Lindą - jedynie 3 kroki!



Mały Świat – przykład 2

Moja znajomość z królem Karolem - jedynie 3 kroki!



Model Małego Świata Model Watts'a-Strogatz'a

Motywacje

Rzeczywiste sieci posiadają relatywnie wysoki współczynnik gronowania i małą średnią długość najkrótszych ścieżek.

Model Małego Świata

Model Watts'a-Strogatz'a

Motywacje

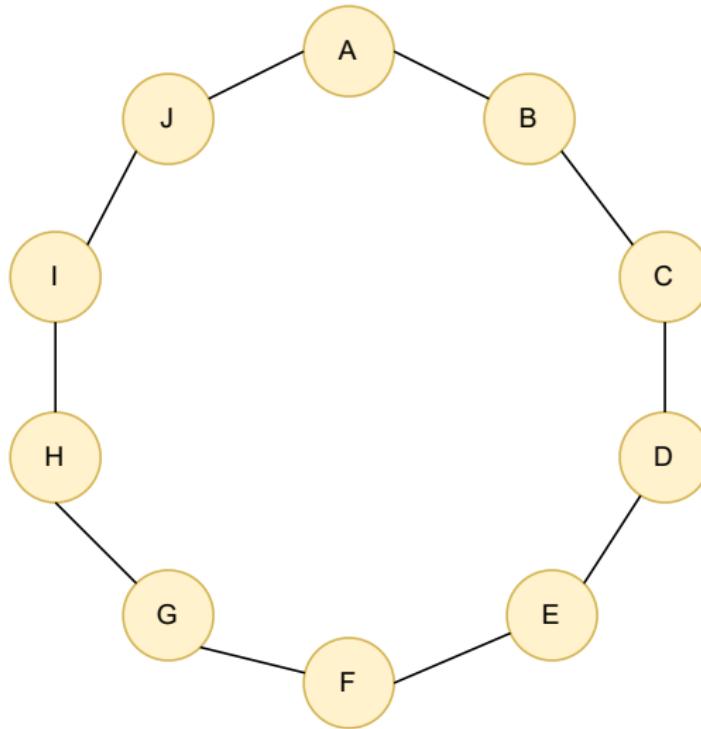
Rzeczywiste sieci posiadają relatywnie wysoki współczynnik gronowania i małą średnią długość najkrótszych ścieżek.

Algorytm budowy

1. Rozpocznij od pierścienia n wierzchołków, w którym każdy wierzchołek jest połączony z k najbliższymi sąsiadami.
2. Ustal parametr $p \in [0, 1]$.
3. Rozważ każdą krawędź (u, v) . Losowo wybierz wierzchołek w i z prawdopodobieństwem p zamień krawędź (u, v) na (u, w) .

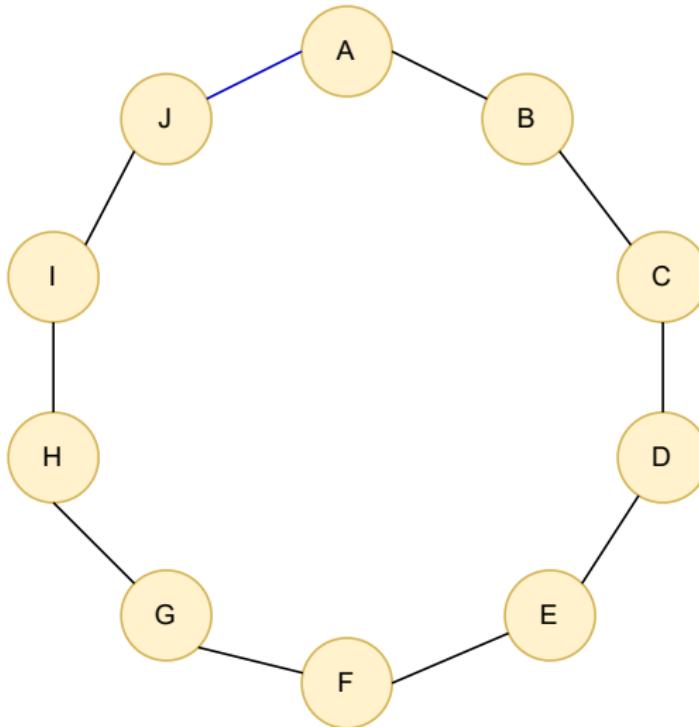
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



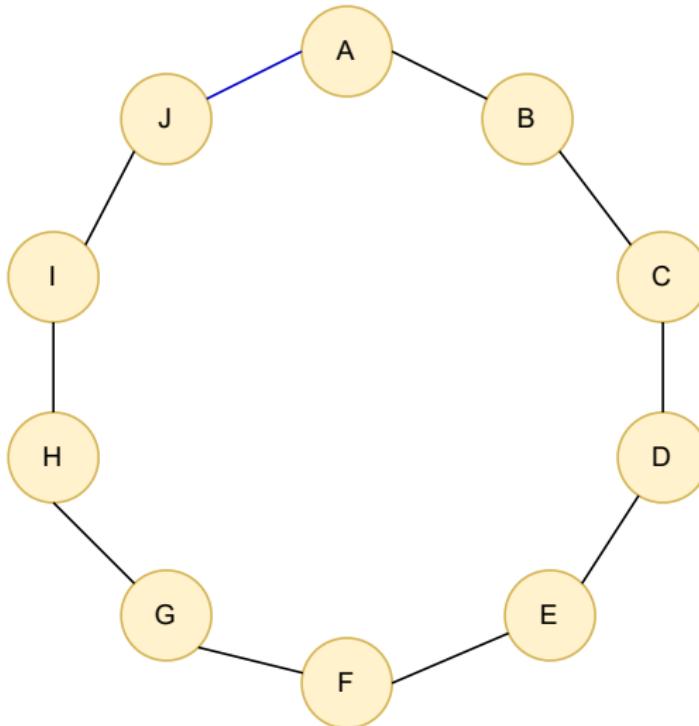
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

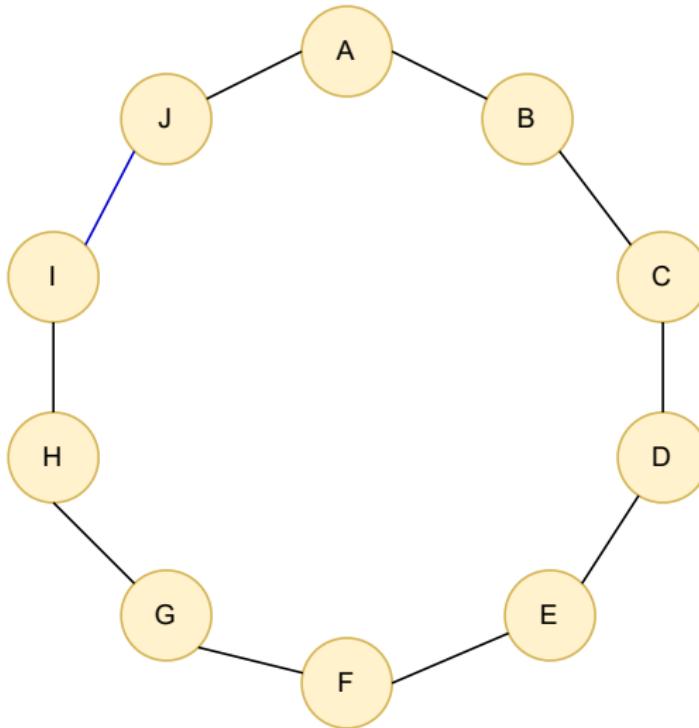
$n = 10, k = 2, p = 0.4$



Nie zamieniamy!

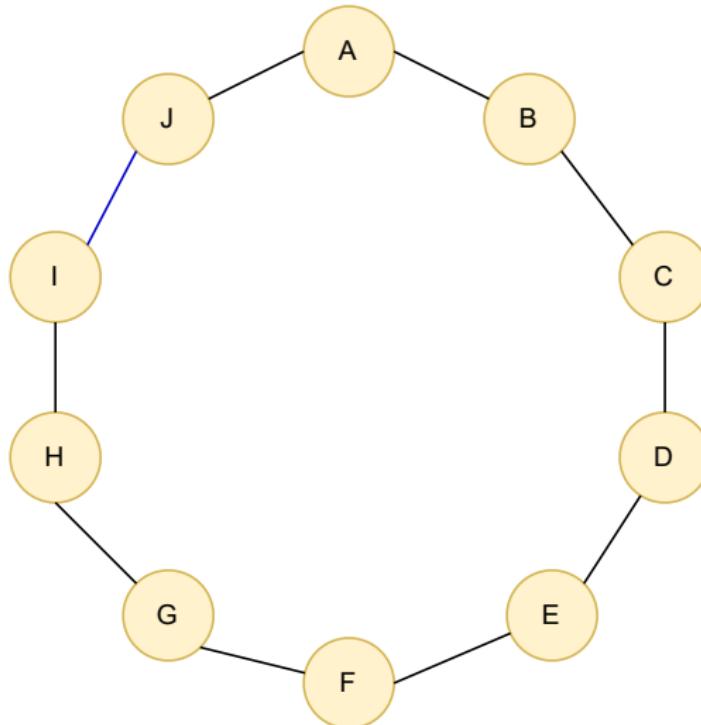
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

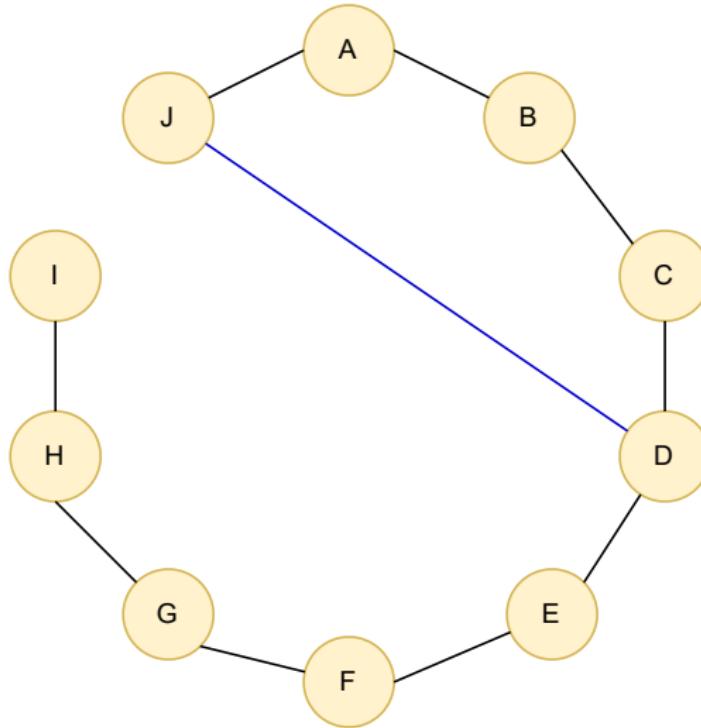
$n = 10, k = 2, p = 0.4$



Zamieniamy!

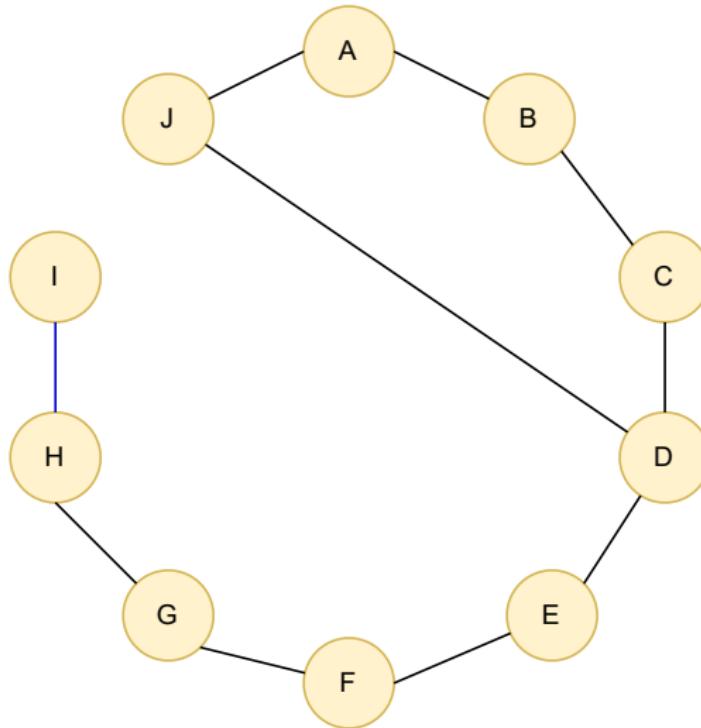
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



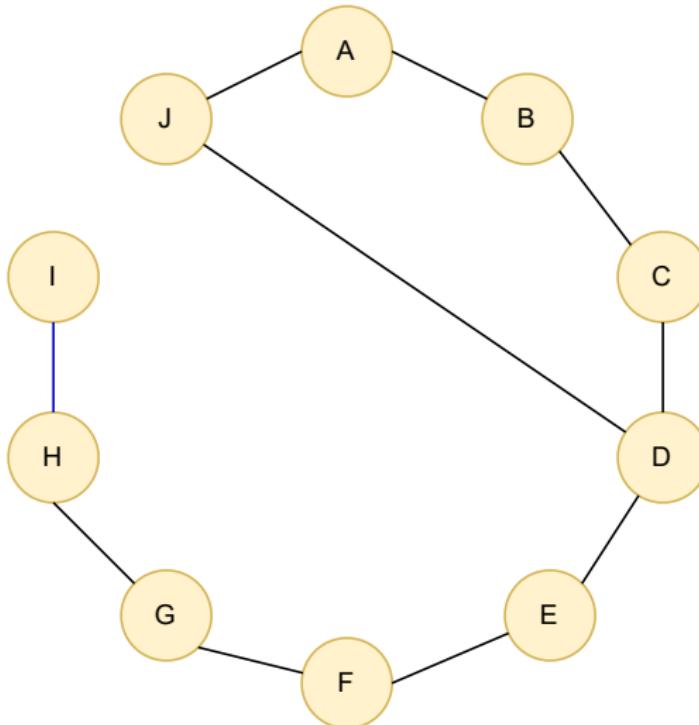
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

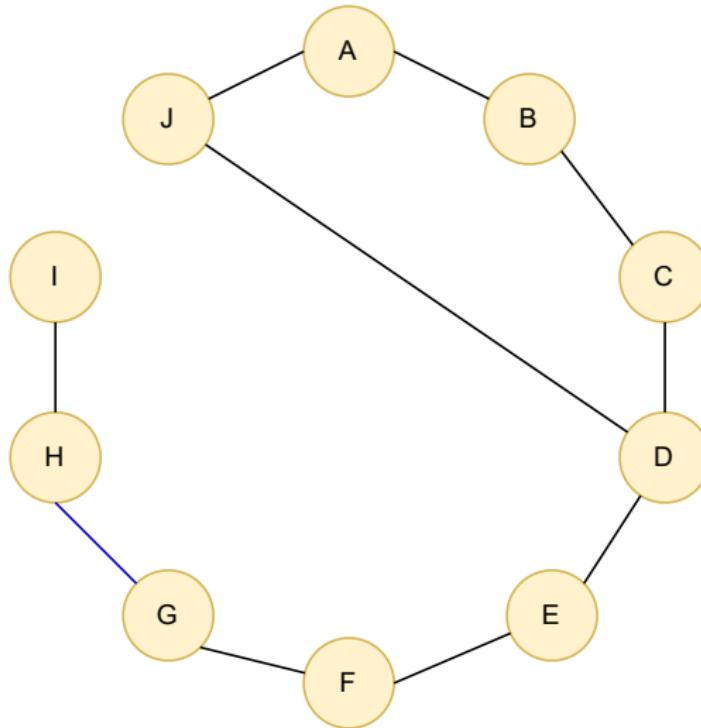
$n = 10, k = 2, p = 0.4$



Nie zamieniamy!

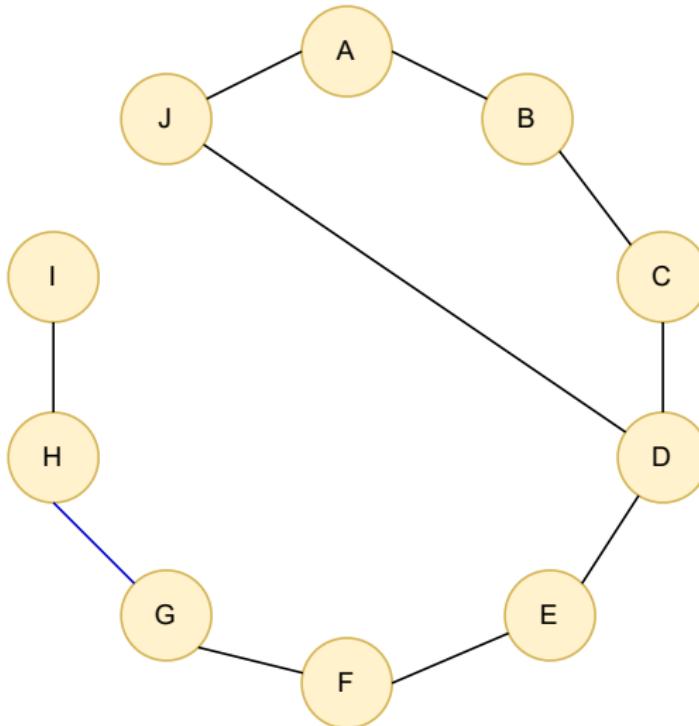
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

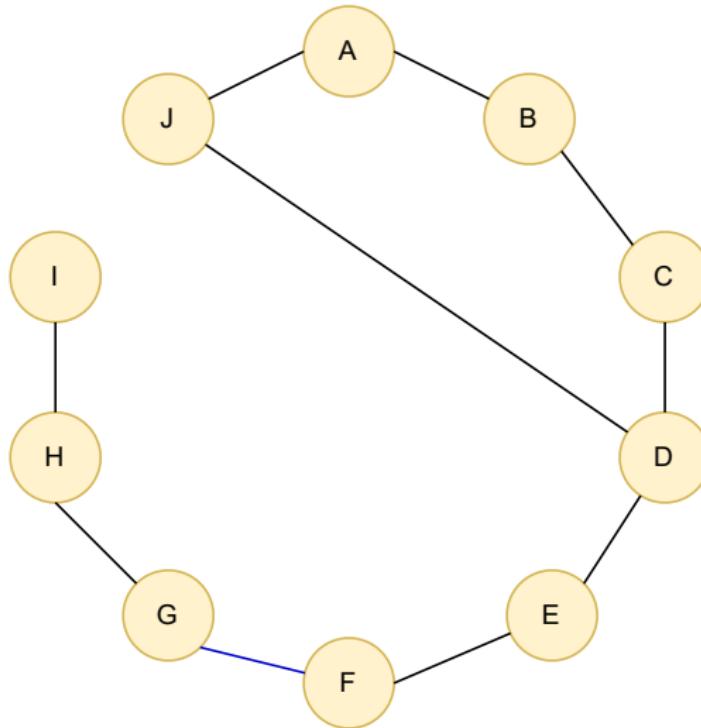
$n = 10, k = 2, p = 0.4$



Nie zamieniamy!

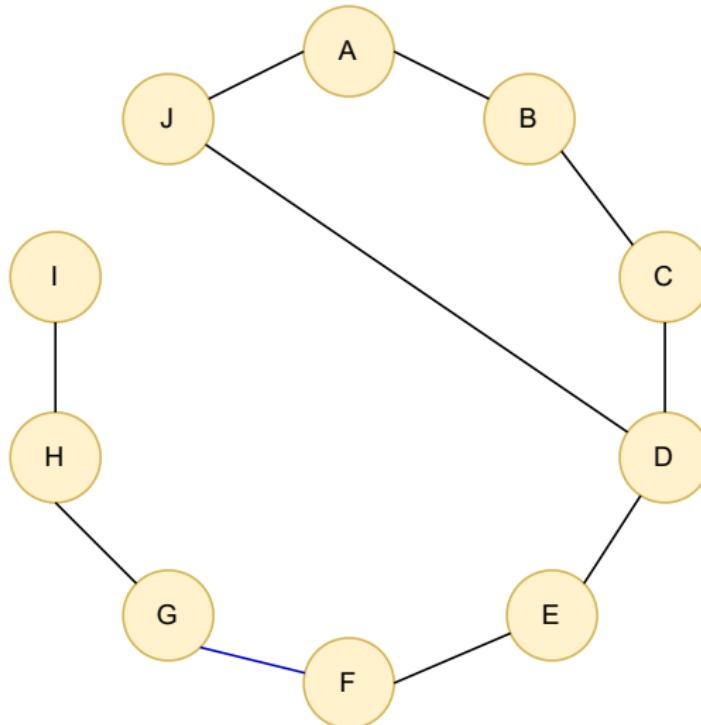
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

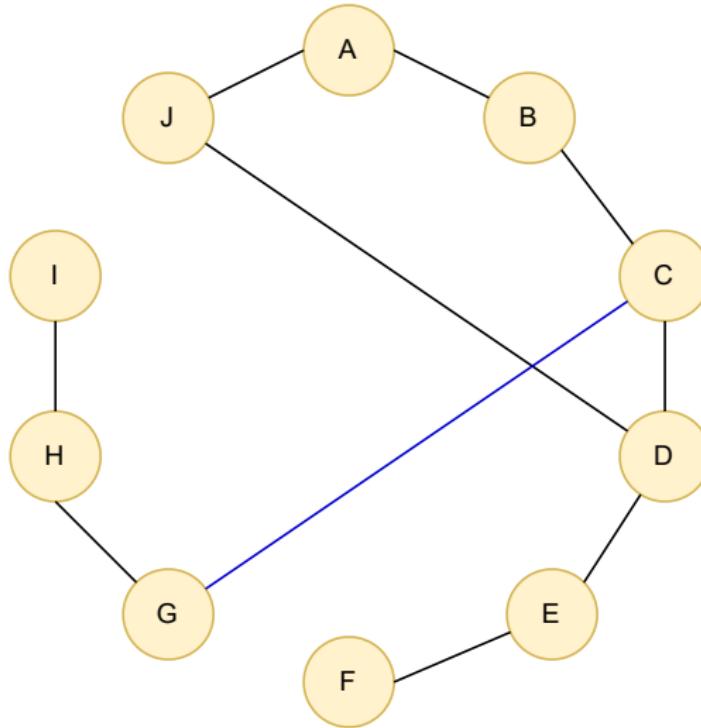
$n = 10, k = 2, p = 0.4$



Zamieniamy!

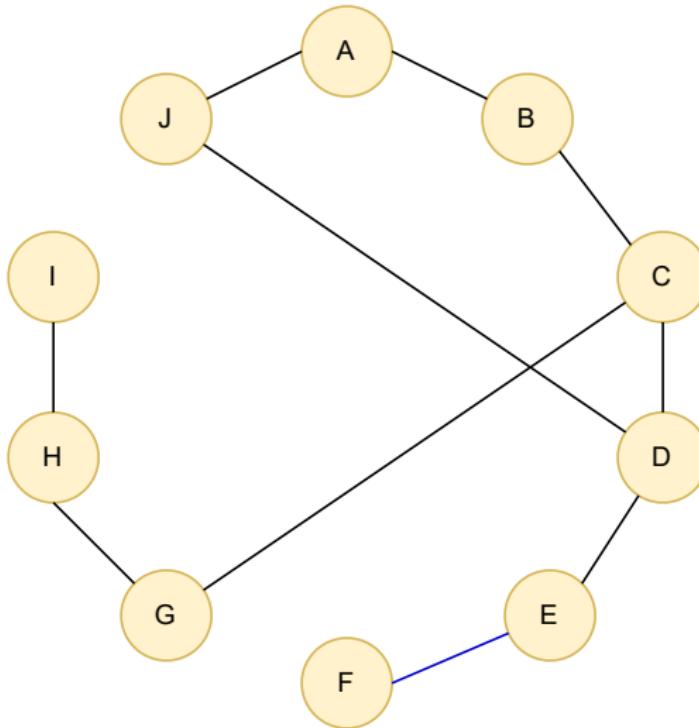
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



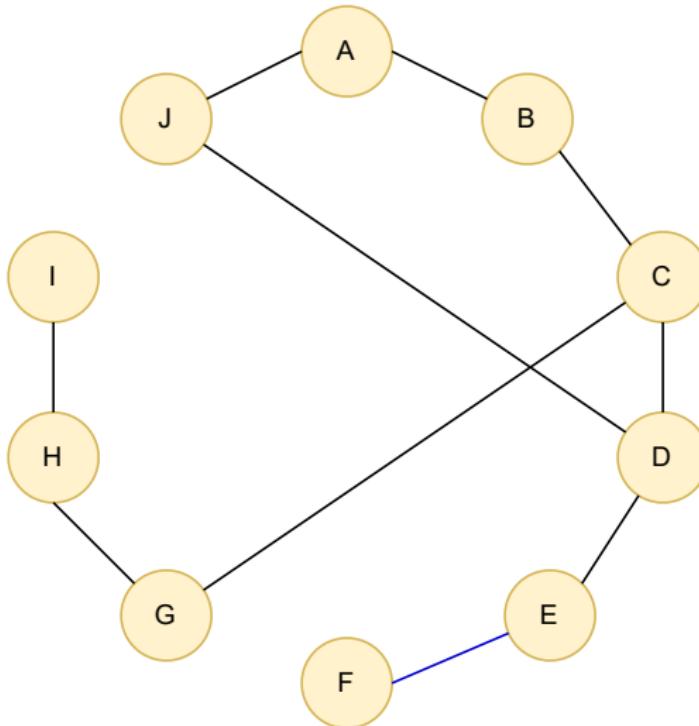
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

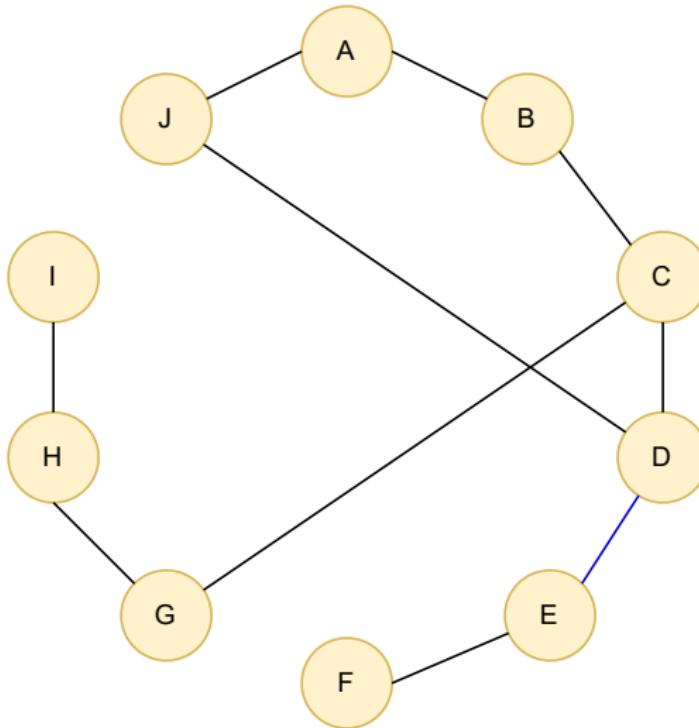
$n = 10, k = 2, p = 0.4$



Nie zamieniamy!

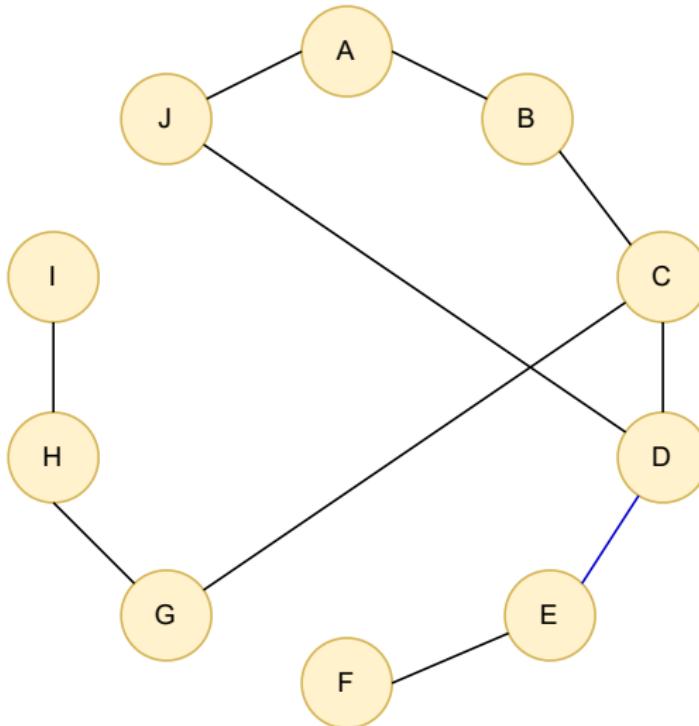
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

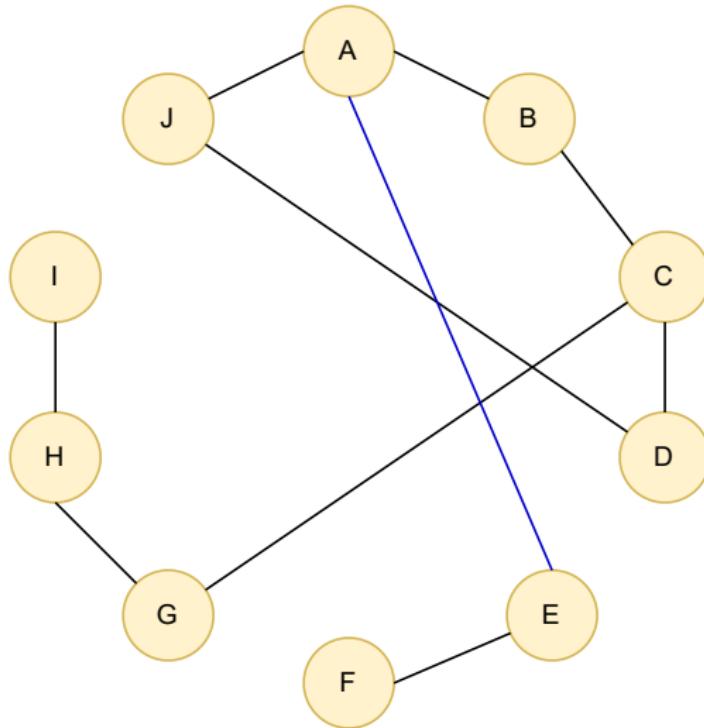
$n = 10, k = 2, p = 0.4$



Zamieniamy!

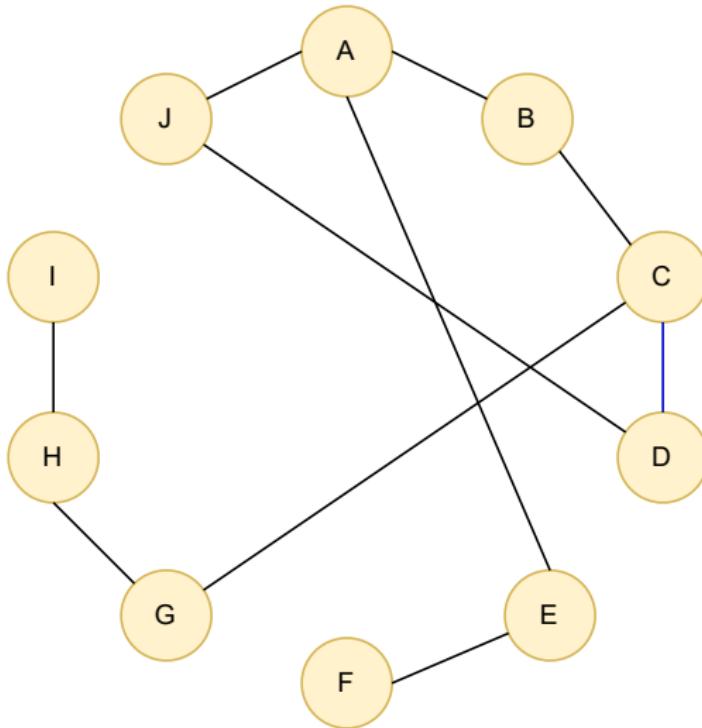
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



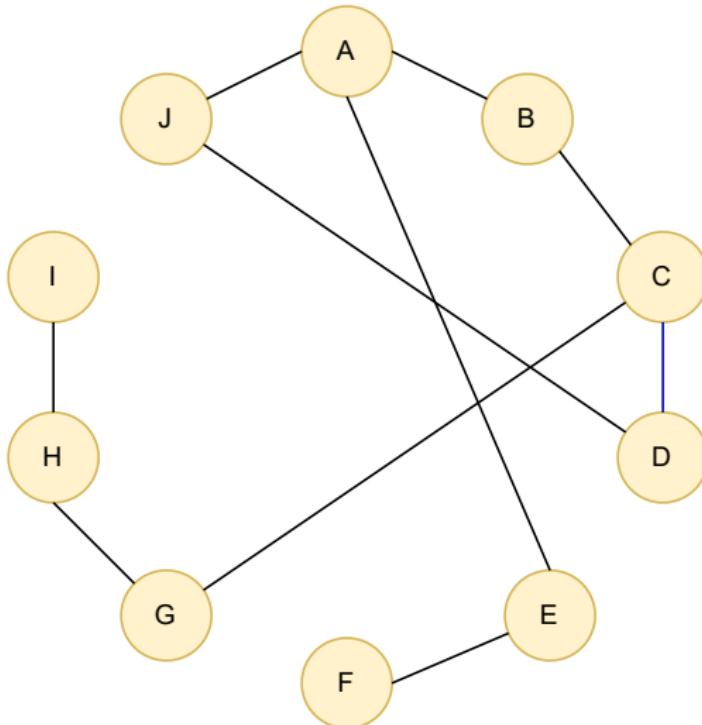
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

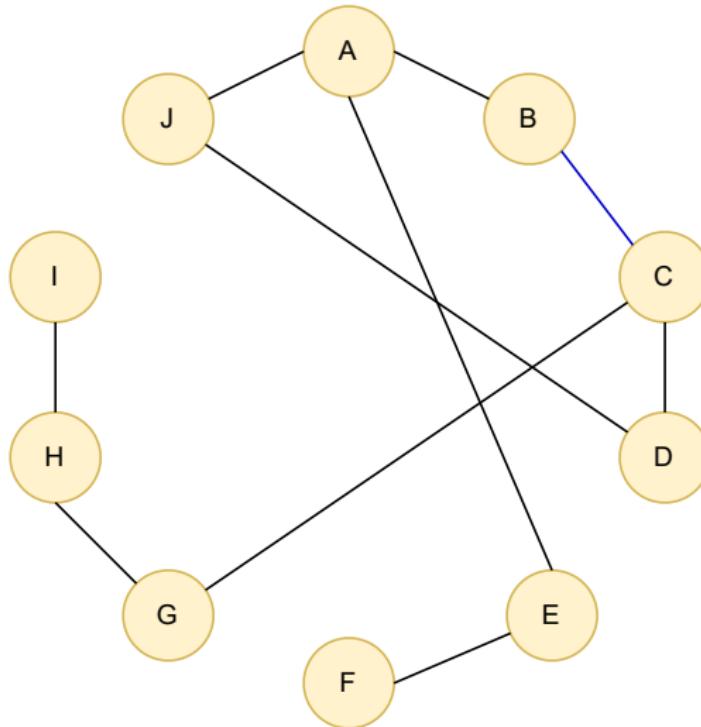
$n = 10, k = 2, p = 0.4$



Nie zamieniamy!

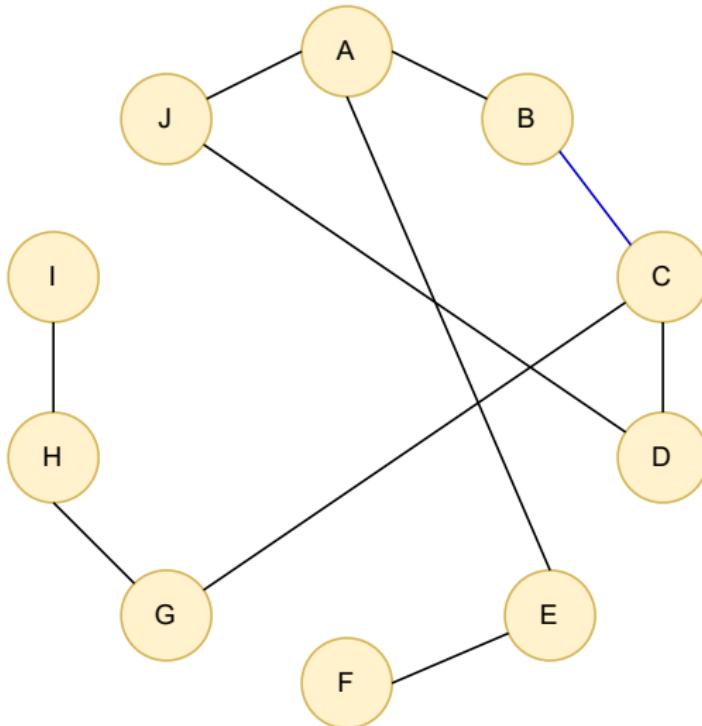
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

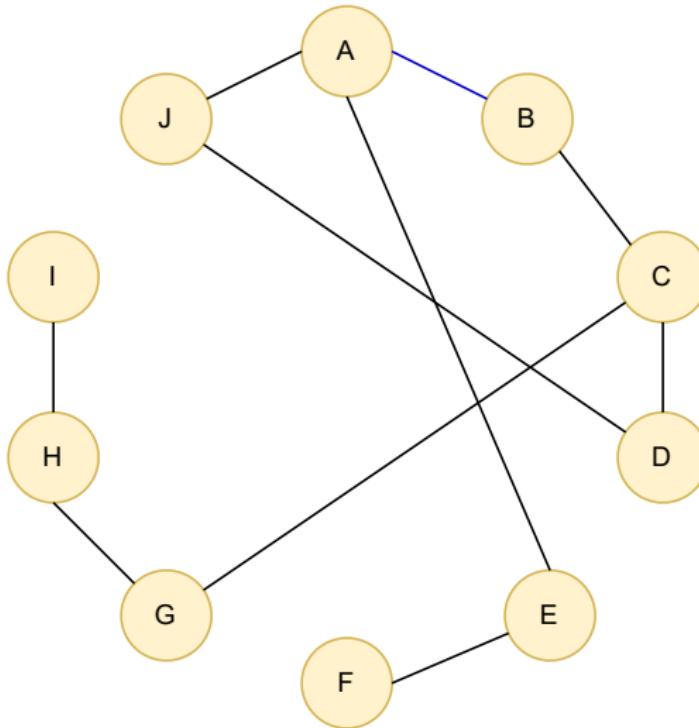
$n = 10, k = 2, p = 0.4$



Nie zamieniamy!

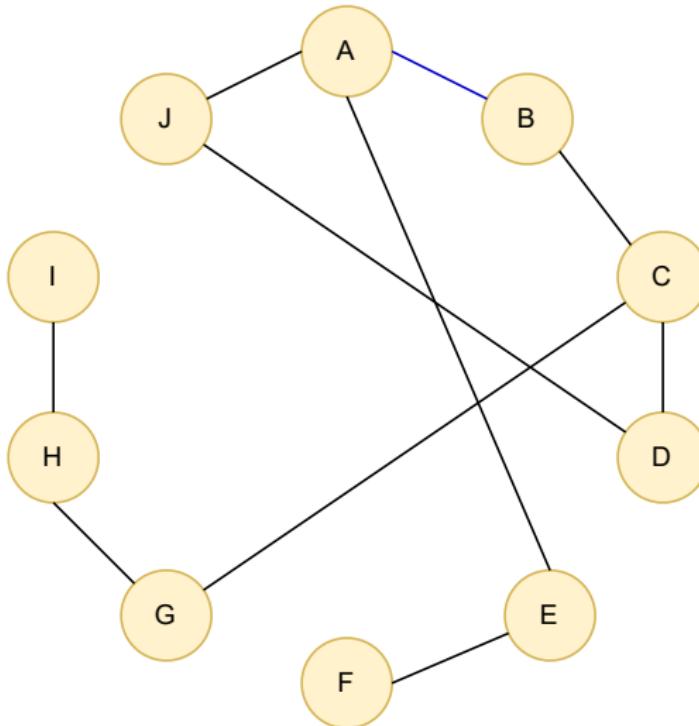
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – przykład

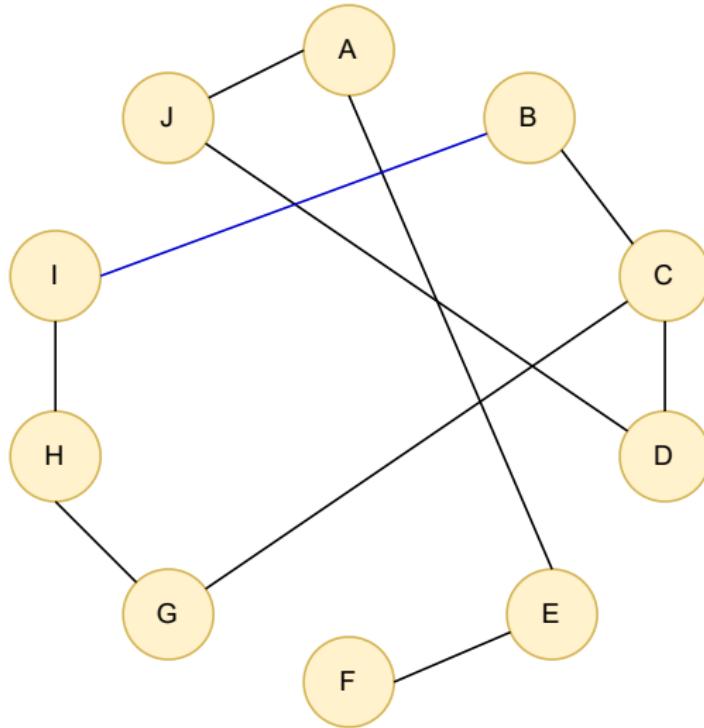
$n = 10, k = 2, p = 0.4$



Zamieniamy!

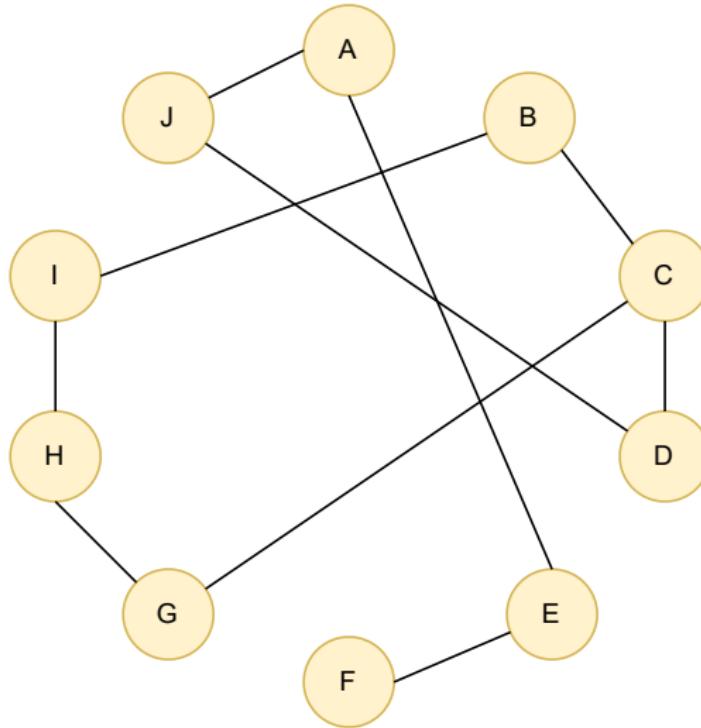
Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$

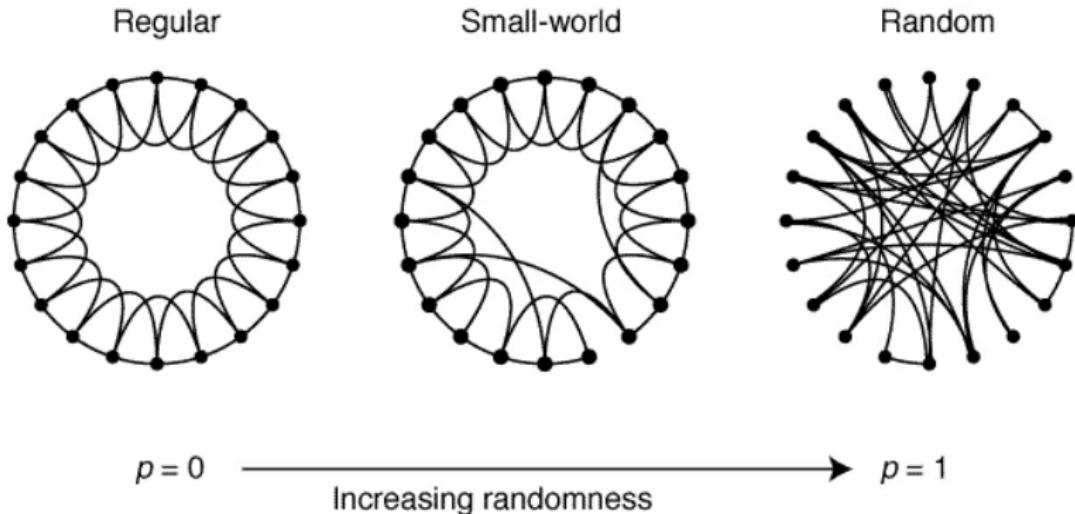


Model Małego Świata – przykład

$n = 10, k = 2, p = 0.4$



Model Małego Świata – wpływ parametru p



Rysunek: Rysunek zaczerpnięty z (Watts i Strogatz, 1998)

1. Daniel Romero, *Applied Social Network Analysis in Python*, University of Michigan, Coursera, 2020.
2. John Scott, *Social Network Analysis*, 4th edition, Sage Publications, 2017.
3. Lawrence Page, Sergey Brin, Rajeev Motwani i Terry Winograd. *The PageRank Citation Ranking : Bringing Order to the Web*, Technical Report, Stanford University, 1998.
4. Watts, D., Strogatz, S. *Collective dynamics of ‘small-world’ networks*. Nature 393, 440–442 (1998). <https://doi.org/10.1038/30918>

Dziękuję

dr inż. Aleksandra Karpus



Fundusze
Europejskie
Polska Cyfrowa



Rzeczpospolita
Polska



Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego

Program Operacyjny Polska Cyfrowa na lata 2014-2020.

Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.

Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Analiza szeregów czasowych



Źródło rysunku: <https://comparic.pl/>

Agnieszka Landowska

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańskia

nailie@pg.edu.pl

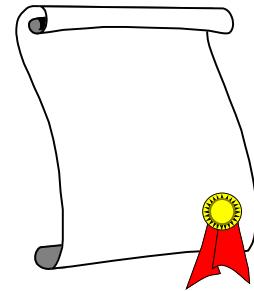
Materiały pomocnicze do wykładu
z Przetwarzanie danych w biznesie na WETI PG.
Ich lektura nie zastępuje wysłuchania wykładu.

Zakres modułu

- Czym jest szereg czasowy
- Składowe szeregu czasowego (składowa stała, trend, okresowość/ sezonowość)
- Wybrane metody analityczne
- Szereg czasowy a prognozowanie
- Przykłady zadań analitycznych

1. Notowania spółki Tesla – 3 lata





Wybrana literatura

Podstawy:

- Avril Coghlan, A Little Book of R For Time Series, Release 0.2, 2016,
<https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf>
- Robert Nau, Principles and risks of forecasting, Fuqua School of Business, Duke University, September 2014,
https://people.duke.edu/~rnau/Principles_and_risks_of_forecasting--Robert_Nau.pdf

R:

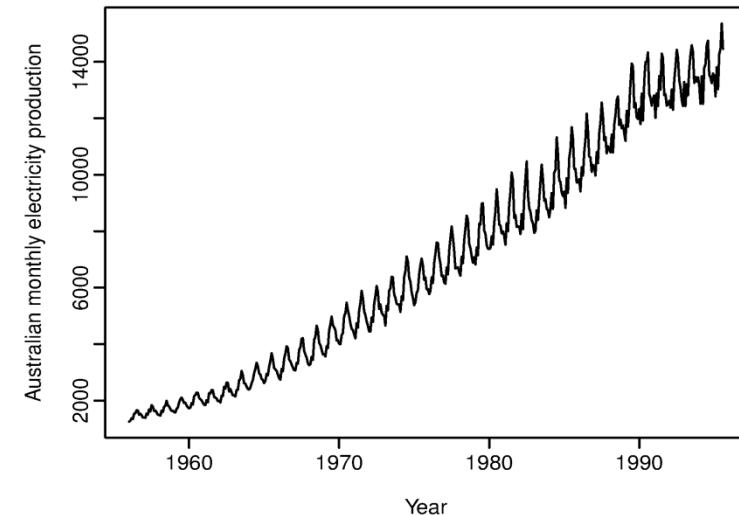
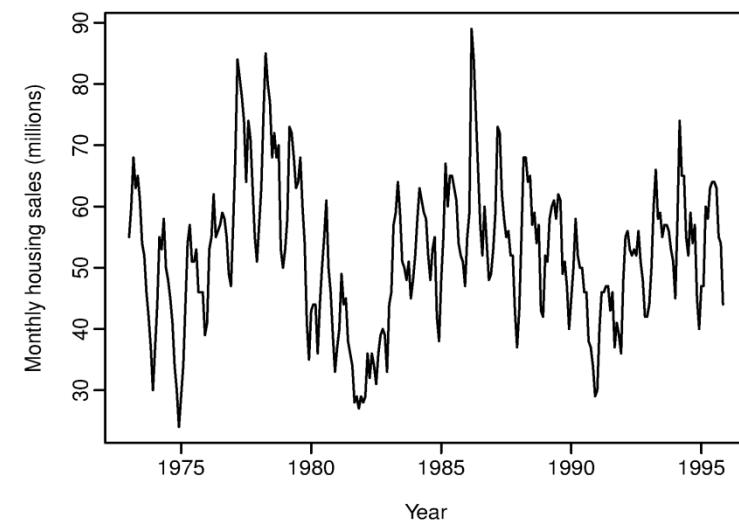
- Vito Ricci, R functions for time series analysis by R.0.5 26/11/04, <https://cran.r-project.org/doc/contrib/Ricci-refcard-ts.pdf>
- Robert H. Shumway, David S. Stoffer, Time Series Analysis and Applications. Using the R Statistical Package, 2016, <http://www.stat.pitt.edu/stoffer/tsa4/>
- CRAN-R project: <https://cran.r-project.org/web/views/TimeSeries.html>

Tutoriale:

- Ryan Womack, Rutgers University, R time series tutorials (Youtube).
 - First part: <https://www.youtube.com/watch?v=QHsmAM6nktY>
- Adam Check: Time Series Forecasting Example in RStudio:
https://www.youtube.com/watch?v=dBNy_A6Zpcc

Szereg czasowy

- Szereg czasowy - ciąg obserwacji pewnego zjawiska w kolejnych jednostkach czasu
- Szereg czasowy *momentów* - szereg zawierający informacje o poziomach badanego zjawiska w określonych momentach pewnego przedziału czasowego
- Szereg czasowy *okresów* - zawiera informacje o rozmiarach zjawiska w ciągu kolejnych okresów danego przedziału czasowego
- ang. *Time-Series Data, Time-related data*
- co różni szereg czasowy od sekwencji danych?
- co różni szereg czasowy od dyskretnego sygnału?





Kroki analizy

- Krok 0. Określ cel analizy
- Krok 1. Poznaj dane
- Krok 2. Analiza podstawowa/wstępna
- Krok 3. Poradź sobie z anomaliami
- Krok 4. Rozłóż szereg czasowy na czynniki pierwsze
- Krok 5. Analiza właściwa (zgodna z celem)
- Krok 6. Ocena wyników i oszacowanie ryzyka



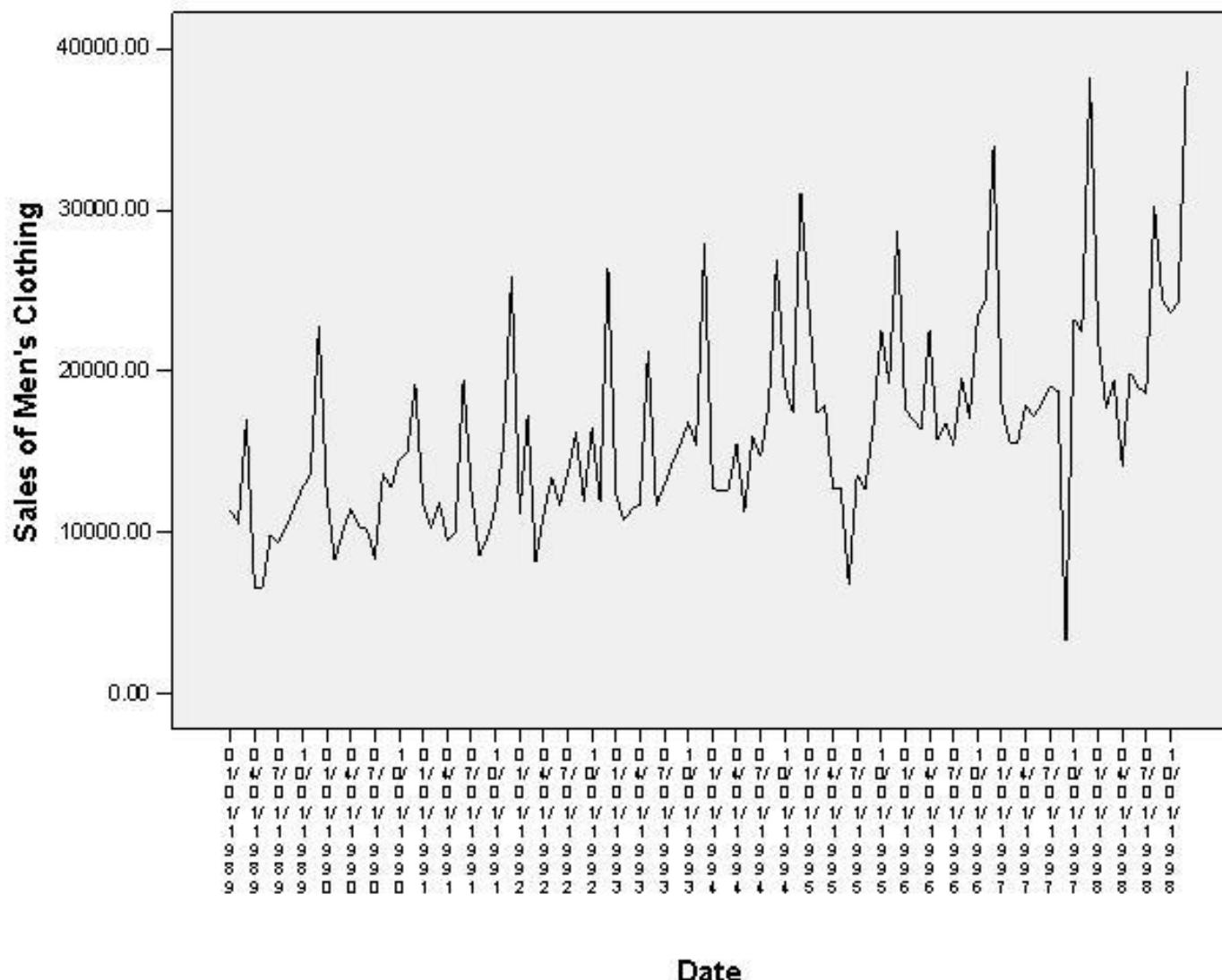
KROK 0. OKREŚL CEL ANALIZY

Krok 0. Jaki jest cel analizy?



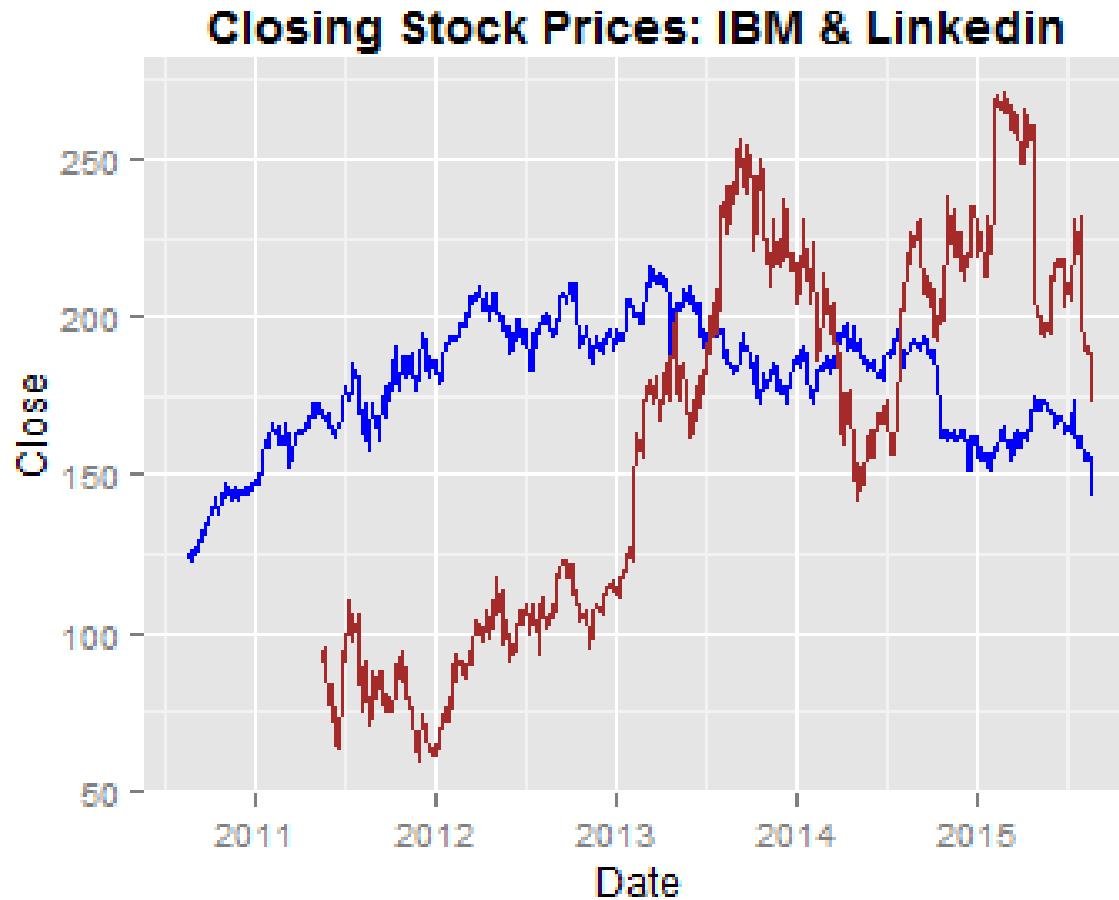
- Badanie występujących w danych: regularnych cykli, wzorców, trendów
- Prognozowanie wartości szeregów dla przyszłych okresów, na podstawie obserwacji historycznych
- Znalezienie modelu dobrze opisującego przebieg danego zjawiska w czasie
- Porównanie zjawiska z dwóch obserwacji – poszukiwanie wzorców, podobieństw
- Poszukiwanie/wykrywanie anomalii

Szereg czasowy – dane sprzedaży



Sprzedaż
ubrań
męskich -
kwartalnie

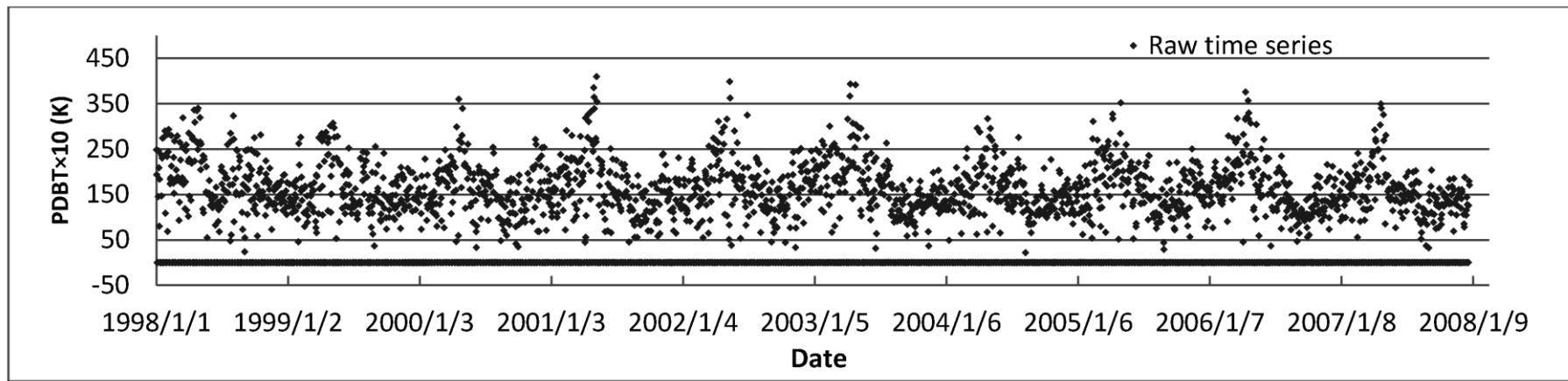
Szereg czasowy – dane giełdowe



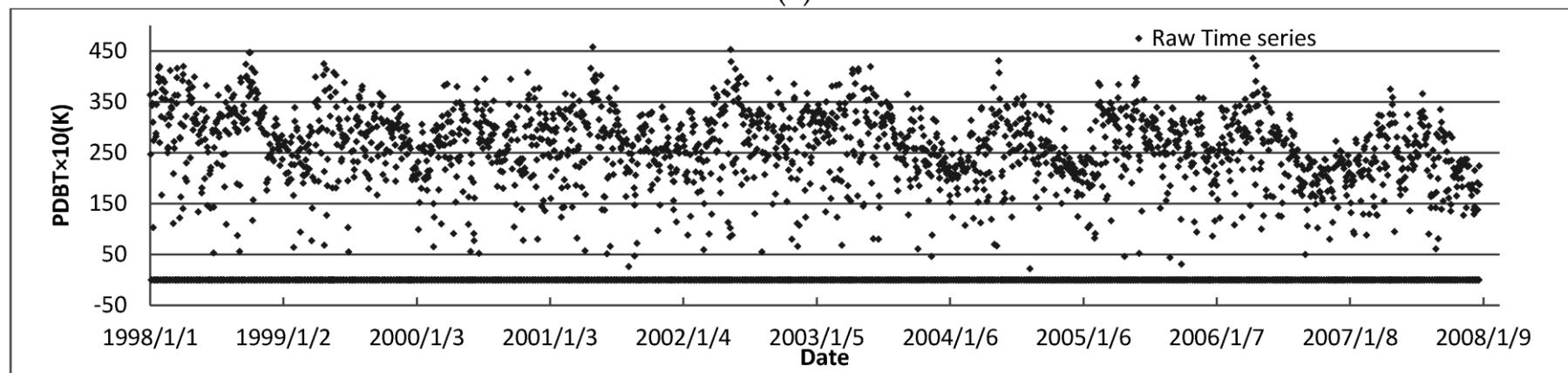
Akcje giełdowe –
ceny zamknięcia

Szereg czasowy – dane pomiarowe

Polarization Difference Brightness Temperature – dane z dwóch czujników w dwóch lokalizacjach



(a)

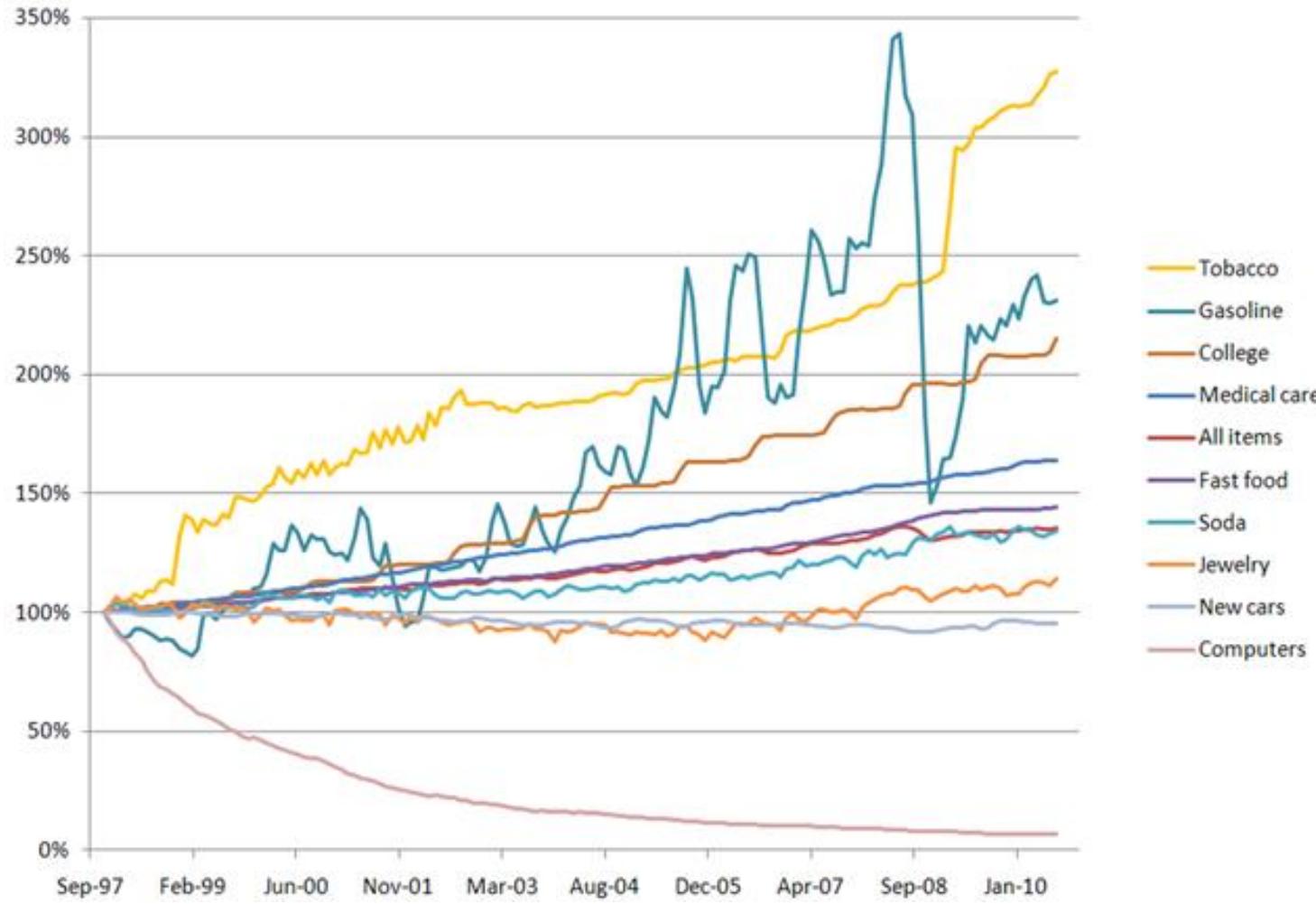


(b)

<http://www.mdpi.com/2072-4292/8/11/970>

Szereg czasowy – zmiany cen

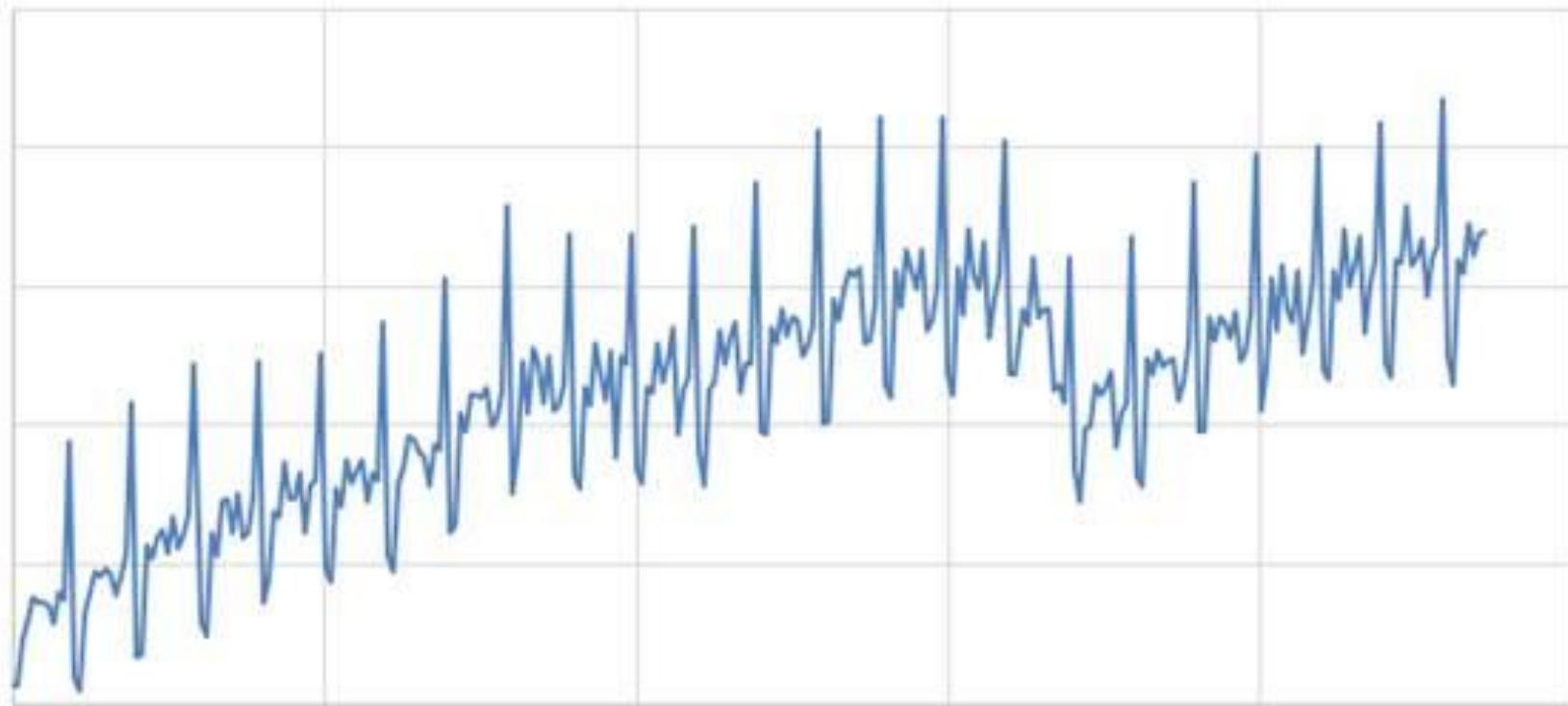
Rynek USA –
zmiany
cen



Szereg czasowy – emocje



Czego to wykres?



Czego to wykres?



Szereg czasowy





KROK 1. POZNAJ DANE

Krok 1. Zaznajom się z danymi

- Skąd pochodzą dane? Jak były zbierane? Od kogo? W jakich warunkach?
Z jaką częstotliwością?
 - Momentów czy okresów?
 - Mierzone czy raportowane?
 - Wiarygodność i stałość procesu pomiaru?
- Jaka jest jednostka danych?
 - Uwaga na metryki stosunkowe i procentowe! Uwaga na jednostki pieniężne w latach! Uwaga na pomiar zmiany!
 - Jaka jest dokładność pomiaru?
- Czy są to dane „surowe” czy poddane już przetworzeniu? Jak były przetwarzane?
- Czy dane są „czyste” czy „brudne”?
 - Czy dane zawierają błędy? Brakujące pomiary? Były mierzone w różnych okresach czasowych? Były zmiany procedury pomiarowej?
- Zobacz dane – wygeneruj wykres (punktowy, liniowy)
 - Czy wszystko się zgadza?
 - Czego można się spodziewać?

Szeregi czasowe w R - podstawy

- Klasy do przechowywania szeregów czasowych:
 - ts (time series)
 - Inne: xts (extensible time series), mts (multiple time series), msts (multi-seasonal time series), its (irregularly spaced time series)
 - class(ts) – sprawdzenie klasy obiektu
- Pakiety główne: stats, forecast, tseries, fpp2
 - library(„package_name”) – import biblioteki
 - library(help = "stats")
- Pakiety dodatkowe: ttr, ast, lmtest, zoo, xts, fts, timeSeries, tis, tframe, MAPA,

Szeregi czasowe w R - podstawy

- **x<-read.ts(...)** – wczytanie szeregu czasowego z pliku (zwraca klasę ts)
- **ts<-ts(x, start=, end=, frequency=)** - przekształcenie dowolnego wektora w serię czasową (do klasy ts)
- **cycle(ts)**: pozycja w cyklu dla obserwacji (stats)
- **deltat(ts)**: zwraca odstęp czasowy między obserwacjami (stats)
- **start(ts)/end(ts)**: zwraca pierwszą/ostatnią obserwację w szeregu (stats)
- **head(ts)/tail(ts)**: początkowe/końcowe obserwacje z szeregu (stats)
- **frequency(ts)**: zwraca częstotliwość próbkowania w jednostce czasu (stats)
- **time(ts)**: zwraca wektor czasów obserwacji (stats)

- **window(ts, start=, end=)** - wybór podzbioru szeregu, także możliwość zmiany próbkowania

Szeregi czasowe w R - wykresy

- **plot(ts), monthplot(ts); seasonplot(ts)**
- **lag.plot(ts)**: plots time series against lagged versions of themselves. Helps visualizing "auto-dependence" even when auto-correlations vanish (stats)
- **monthplot(ts)**: plots a seasonal (or other) subseries of a time series (stats)
- **plot.ts(ts)**: plotting time-series objects (stats)
- **seaplot(ts)**: plotting seasonal sub-series or profile (ast)
- **seqplot.ts(ts)**: plots a two time series on the same plot frame (tseries)
- **tsdiag(ts)**: a generic function to plot time-series diagnostics (stats)
- **ts.plot(ts)**: plots several time series on a common plot. Unlike 'plot.ts' the series can have a different time bases, but they should have the same frequency (stats)



KROK 2. ANALIZA PODSTAWOWA

Krok 2. Szereg czasowy – analiza podstawowa

Analiza składowej stałej

- **Analiza tendencji centralnej**
 - średnia arytmetyczna albo chronologiczna
 - mediana
- **Analiza zmienności**
 - wariancja, odchylenie standardowe,
 - wartości minimalne i maksymalne, rozrzut,
 - kwartyle, odstęp Q3-Q1

dla szeregów momentów

Analiza dynamiki

- **Przyrosty**
 - absolutne lub względne
 - jednopośrednie lub łańcuchowe
- **Indeksy dynamiki**
 - jednopośrednie
 - łańcuchowe
- **Inne miary**
 - średniookresowe tempo zmian
 - indeksy agregatowe

$$\bar{y} = \frac{\frac{y_1+y_2}{2} + \frac{y_2+y_3}{2} + \dots + \frac{y_{n-1}+y_n}{2}}{n}$$

Analiza dynamiki - wzory

- Przyrost absolutny jednopodstawowy (j - okres bazowy)

$$\Delta x_{n/j} = x_n - x_j$$

- Przyrost absolutny łańcuchowy

$$\Delta x_{n/n-1} = x_n - x_{n-1}$$

- Przyrost względny jednopodstawowy (j - okres bazowy)

$$\dot{\Delta} x_{n/j} = \frac{x_n - x_j}{x_j}$$

- Przyrost względny łańcuchowy

$$\dot{\Delta} x_{n/n-1} = \frac{x_n - x_{n-1}}{x_{n-1}}$$

- Indeks jednopodstawowy (j - okres bazowy)

$$x_{n/j} = \frac{x_n}{x_j}$$

- Indeks łańcuchowy

$$x_{n/n-1} = \frac{x_n}{x_{n-1}}$$

Analiza dynamiki – przykład (1)

Szereg czasowy okresów:

| lata | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|------------------------|-------|-------|-------|-------|-------|-------|-------|
| urodzenia | | | | | | | |
| żywego w Polsce w tys. | 378,3 | 368,2 | 353,8 | 351,1 | 356,1 | 364,4 | 374,2 |

Źródło: Roczniki Demograficzne.

przyrosty absolutne i względne

| okresy czasu (lata) | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|------|-------|-------|-------|-------|-------|------|
| przyrosty absolutne w tys. (rok bazowy - 2000) | 0 | -10,1 | -24,5 | -27,2 | -22,2 | -13,9 | -4,1 |
| przyrosty względne w % (rok bazowy - 2000) | 0 | -2,7 | -6,5 | -7,2 | -5,9 | -3,7 | -1,1 |
| przyrosty absolutne w tys. (rok bazowy - 2003) | 27,2 | 17,1 | 2,7 | 0 | 5 | 13,3 | 23,1 |
| przyrosty względne w % (rok bazowy - 2003) | 7,8 | 4,9 | 0,8 | 0 | 1,4 | 3,8 | 6,6 |

Źródło przykładu:

http://www.demografia.uni.lodz.pl/dlastud/szeregi_czasowe_I.pdf

Analiza dynamiki – przykład (2)

Szereg czasowy okresów:

| lata | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|----------------------|-------|-------|-------|-------|-------|-------|-------|
| urodzenia | | | | | | | |
| żywe w Polsce w tys. | 378,3 | 368,2 | 353,8 | 351,1 | 356,1 | 364,4 | 374,2 |

Źródło: Roczniki Demograficzne.

indeksy

| lata | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|------|------|------|------|-------|-------|-------|
| indeksy łańcuchowe w % (rok poprzedni=100) | – | 97,3 | 96,1 | 99,2 | 101,4 | 102,3 | 102,7 |
| indeksy jednopodstawowe w % (rok 2000=100) | 100 | 97,3 | 93,5 | 92,8 | 94,1 | 96,3 | 98,9 |

Źródło przykładu:

http://www.demografia.uni.lodz.pl/dlastud/szeregi_czasowe_I.pdf



KROK 3. PORADŹ SOBIE Z ANOMALIAMI

Krok 3. Anomalie, wartości odstające i brakujące

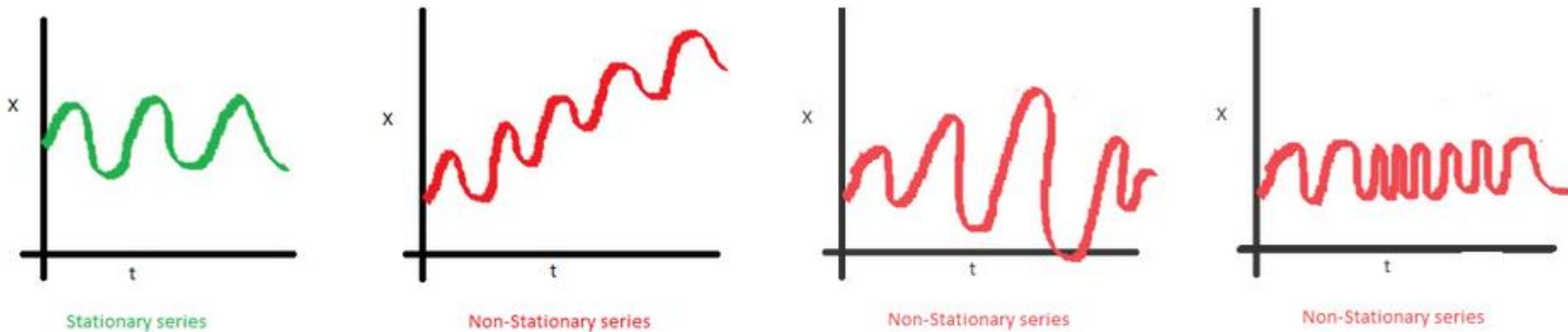
- Obsługa danych brakujących i odstających zależy od celu analizy
- Dane brakujące – czym zastąpić?
 - Zerem/inną wartością neutralną?
 - Średnią?
 - Medianą?
- Wartość odstająca:
 - Poza przedziałem: $\text{mean} \pm 3x \text{ SD}$
 - Poza przedziałem: $\text{median} \pm 3x (\text{Q3}-\text{Q1})$
- Czym zastąpić?
- Podstawowe metryki (średnia, wariancja, przyrosty itp. – krok 2)
 - podajemy bez modyfikacji i po modyfikacji

Szeregi czasowe w R - podstawy

- Statystyki:
 - **mean(ts)**, **sd(ts)**, **median(ts)**, **IQR(ts)**, **qqplot(ts)**, **quantile(ts)**,
weighted.mean(ts)
- Indeksy:
 - **diff.ts(ts,1)**: oblicza różnice (stats)
 - **lag(ts,1)**: wprowadza opóźnienie (stats)
 - **diff(ts1,ts2)** – różnice między dwoma szeregami czasowymi
 - **diff(ts, differences=1)** – różnice między obserwacjami tego samego wykresu
- Wartości brakujące:
 - **na.exclude(ts)**; **na.fail(ts)**; **na.omit(ts)**; **na.omit.ts(ts)**; **na.pass(ts)**;
napredict(ts); **naprint(ts)**; **naresid(ts)**
- Połączenia szeregów:
 - **ts.intersect(ts1, ts2)**; **ts.union(ts1, ts2)**; **ts.plot(ts1, ts2)**

Szeregi stacjonarne i niestacjonarne

- Szeregi (sygnały) stacjonarne – szereg (sygnał), którego charakterystyki statystyczne nie są zmienne w czasie (średnia, wariancja, autokorelacja)
- ang. stationary vs non-stationary
- Statystyczne metody predykcyjne są najczęściej oparte na założeniu stacjonarności (np. ARMA)
- Jakie są najczęściej szeregi czasowe w praktyce?
- Szeregi można próbować sprowadzić do stacjonarnych przez różne przekształcenia
 - Obliczenie różnic (difference-stationary)
 - Odsiewanie trendu (trend-stationary)



Stacjonaryzacja szeregu

- Metoda 1: Logarytmowanie
 - `ts<- log(ts)`
- Metoda 2: Indeksy zmian
 - `diff.ts(ts,1)`: returns suitably lagged and iterated differences (stats)
 - `lag(ts,1)`: computes a lagged version of a time series, shifting the time base back by a given number of observations (stats)
 - `diff(ts, differences=1)`
 - `ndiffs(ts)`: number of differences required to achieve stationarity (forecast)
- Metoda 3: Estymacja trendu i odjęcie od oryginalnego szeregu
- Metoda 4: Estymacja cykli/sezonowości i odjęcie
- Inne metody:
 - zmiana dziedziny analizy (częstotliwości i czasu-częstotliwości), filtrowanie częstotliwości
 - `adf.test(ts)` Augmented Dickey-Fuller test. Rejecting the null hypothesis suggests that a time series is stationary (from the [tseries](#) package)

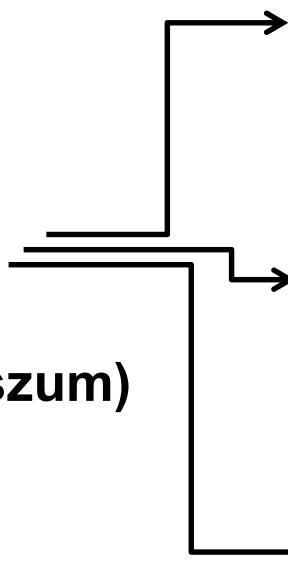


KROK 4. ROZŁÓŻ NA CZYNNIKI PIERWSZE

Krok 4. Dekompozycja szeregu czasowego

Składowe szeregu czasowego

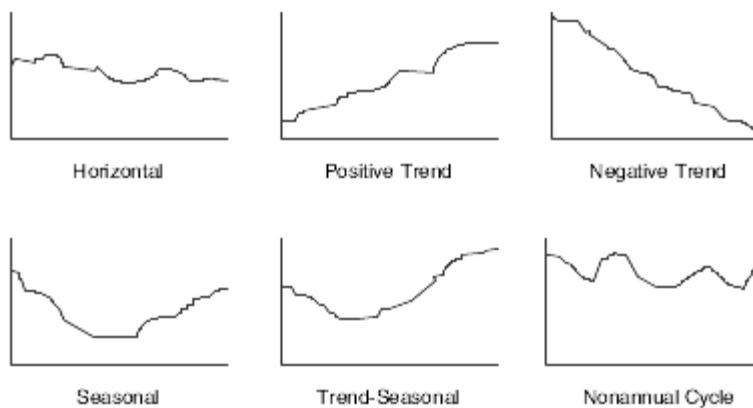
- Część systematyczna
 - Składowa stała
 - Trend
 - Składowa okresowa
- Część przypadkowa (szum)



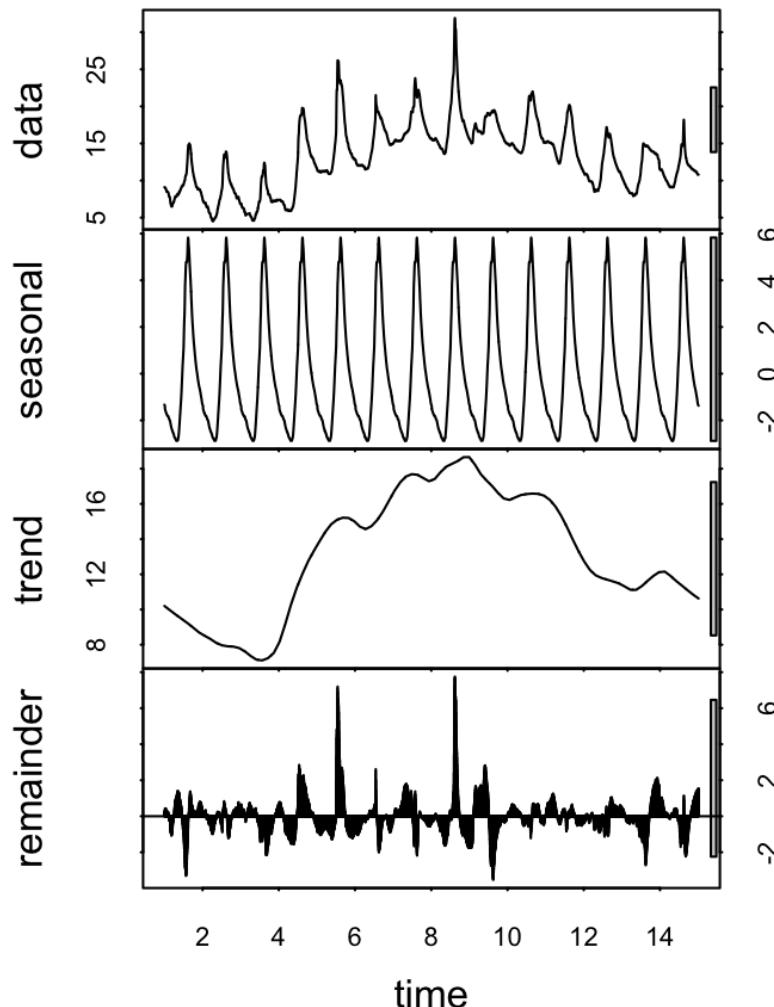
Składowa okresowa

- wahania cykliczne - długookresowe, rytmiczne wahania (cykl koniunkturalny gospodarki, cykl rozwoju populacji nabywców danego produktu, itp.),
- wahania sezonowe – krótkookresowe do 1 roku, odzwierciedlają wpływ zachowań wynikający z „kalendarza” (np. rytm pracy w skali tygodnia, dnia, pory roku, świąt, ...)
- okresowość – krótkoterminowa powtarzalność zmiany amplitudy wynikająca z charakteru badanego zjawiska (ruch kołowy, rytm serca)

Six Typical Demand Patterns



Dekompozycja szeregu



- Różne cele dekompozycji:
 - analiza trendu
 - analiza cykliczności
 - dopasowanie modelu
- Dekompozycję powtarzamy tak długo, aż pozostała część szeregu (remainder, residuals) będzie szumem białym
- White noise:
 - Średnia = 0
 - Wariancja stała w czasie
 - Brak autokorelacji

Źródło rysunku: <https://www.r-bloggers.com/time-series-analysis-with-r-testing-stuff-with-netatmo-data/>

Time Series in R - dekompozycja

- **stl(ts)**: decomposes a time series into seasonal, trend and irregular components (stats)
- **decompose(ts)**: decomposes a time series into seasonal, trend and irregular components using moving averages. Deals with additive or multiplicative seasonal component (stats)
- **filter(ts)**: linear filtering on a time series (stats)
- **HoltWinters(ts)**: computes Holt-Winters Filtering of a given time series (stats)
- **sfilter(ts)**: removes seasonal fluctuation using a simple moving average (ast)
- **spectrum(ts)**: estimates the spectral density of a time series (stats)
- **tsr(ts)**: decomposes a time series into trend, seasonal and irregular. Deals with additive and multiplicative components (ast)

Rodzaje modeli

• Addytywne

Data = Seasonal effect

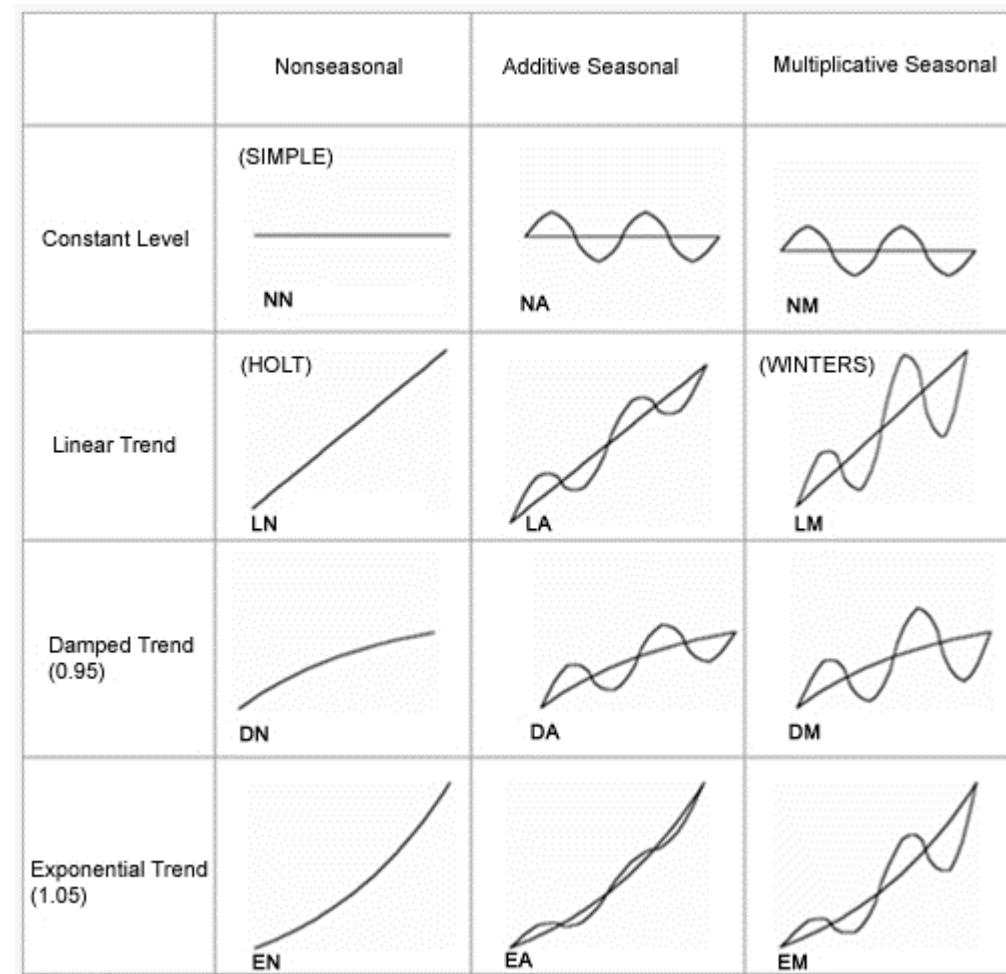
- + Trend
- + Cyclical
- + Residual

• Multiplikatywne

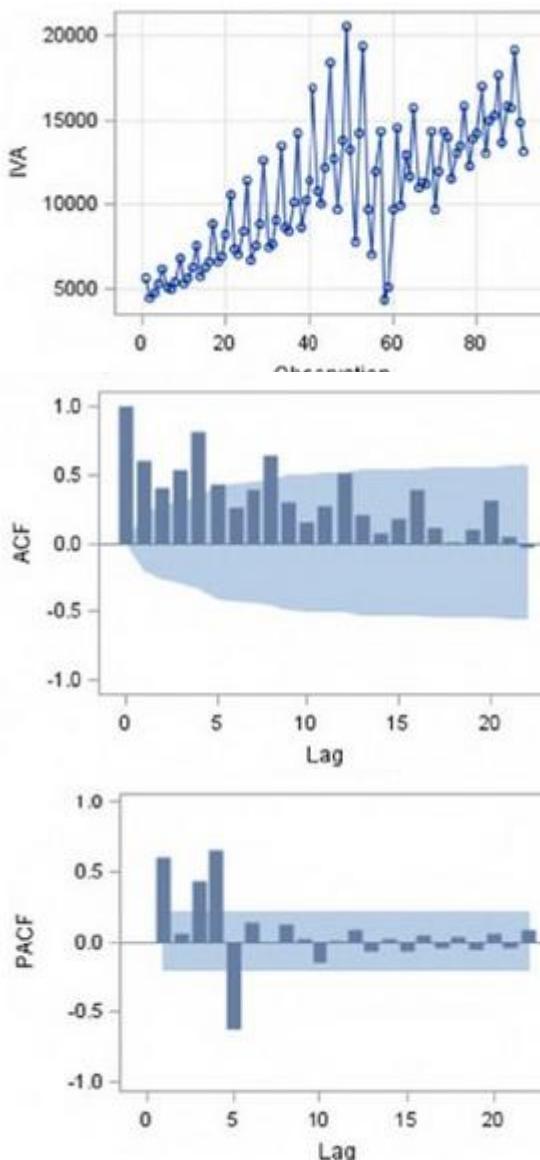
Data = Seasonal effect

- * Trend
- * Cyclical
- * Residual

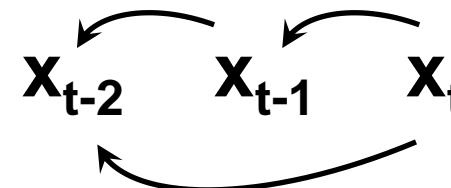
• Mieszane



Autokorelacja



- na ile obserwacja w okresie czasu t zależy od poprzednich obserwacji?
- autokorelacja – korelacja z poprzednimi wartościami



Interpretacja:

- stały wysoki poziom ACF oznacza trend
- peak w poziomie ACD i PACF oznacza okresowość
- może służyć do sprawdzenia, czy usunięto z szeregu składową stałą i okresową



KROK 5. ZASADNICZA ANALIZA (ZGODNA Z CELEM)

Kolejne kroki analizy zależą od jej celu

- Cel: Analiza
 - Analiza trendu
 - Analiza składowych okresowych (sezonowości, cykli)
- Cel: Predykcja
 - Budowa modelu predykcyjnego
 - Ocena jego dokładności
- Cel: Wykrywanie
 - Poszukiwanie wzorców (motif discovery)
 - Detekcja anomalii (novelty detection)
 - Porównywanie szeregów

Modele predykcyjne w R - podstawy

Ogólny schemat przygotowania modelu:

- `model <- metoda_obliczania_modelu(ts, parametry)`
 - Uwaga! `ts` powinien być wcześniej ustacjonaryzowany

Ocena modelu:

- `accuracy(model)` – oblicza dokładność modelu predykcji
- `coefficients(model)` – współczynniki modelu
- `summary(model)` – podsumowanie modelu

Wyświetlanie:

- `forecast(model, 5)`
- `plot(forecast(model,5))`

Analiza i prognoza trendu

- Wygładzanie
 - Średnia ruchoma (krocząca) - polega na zastąpieniu każdego elementu szeregu przez zwykłą lub ważoną średnią n sąsiadujących wartości, gdzie n jest szerokością okna wygładzania
 - Wygładzanie medianą (bardziej odporne na wartości odstające)
 - NKWO – metoda wygładzania (najmniejszych kwadratów ważonych odległościami)
- Dopasowanie funkcji
 - Jeśli trend jest monotoniczny (stale rosnący lub malejący), analiza jest (relatywnie) prosta
 - jeśli występuje wyraźny monotoniczny składnik nieliniowy, dane lepiej najpierw przekształcić w celu usunięcia nieliniowości
 - logarytm, funkcja wykładnicza albo wielomianowa
 - np. autoregresja

Analiza i prognoza składowej okresowej

- Sezonowość – zależność stopnia k pomiędzy obserwacją i oraz $i-k$ (k-lag)
- Mierzona przez autokorelację
 - Krok 1. Wygenerowanie autokorelogramów ACF() i analiza autokorelacji
 - Krok 2. Analiza autokorelacji częściowej PACF()
- Określenie parametrów modeli predykcyjnych
- Usuwanie zależności autokorelacyjnych (różnicowanie)
 - W celu głębszej analizy zjawiska autokorelacji
 - W celu stacjonaryzacji sygnału



Analiza trendu i okresowości w R (1)

- **SMA()**: simple moving average (ttr)
- **acf(), pacf(), ccf()**: the function 'acf' computes (and by default plots) estimates of the autocovariance or autocorrelation function. Function 'pacf' is the function used for the partial autocorrelations. Function 'ccf' computes the cross-correlation or cross-covariance of two univariate series (stats)
- **ar()**: fits an autoregressive time series model to the data, by default selecting the complexity by AIC (stats)

Analiza trendu i okresowości w R (2)

ARMA/ARIMA

- ARIMA – połączenie modelu średniej kroczącej i modelu autoregresyjnego
- Warunek wejściowy - stacjonarność szeregu
 - Sprawdzenie autokorelacji (~ 0)
 - Testy statystyczne:
 - Test Ljung-Box `Box.test(ts, type="Ljung-Box")`
 - Augmented Dickey-Fuller Test `adf.test(ts)`
 - Kwiatkowski-Phillips-Schmidt-Shin (KPSS) `kpss.test(ts, null="Trend")`
- Dobór parametrów
 - `Arima(ts, order=c(p,d,q))`: fits an ARIMA model to a univariate time series (stats)
 - `Arma()`: fits an ARMA model to a univariate time series by conditional least squares (tseries)
- Automatyczny dobór parametrów
 - `auto.arima (ts)`

Analiza trendu i okresowości w R (3)

modele eksponentjalne

- ets(ts)
- Model Holt-Winters:
 - simple exponential - models level
`fit <- HoltWinters(ts, beta=FALSE, gamma=FALSE)`
 - double exponential - models level and trend
`fit <- HoltWinters(ts, gamma=FALSE)`
 - triple exponential - models level, trend, and seasonal components
`fit <- HoltWinters(ts)`



KROK 6. OCEŃ WYNIKI I RYZYKO

Ocena dopasowania modelu

- ocena residuals – pozostałości po odjęciu modelu od wartości szeregu
 - bliskość do białego szumu
- minimalizacja błędu prognozy *ex post*

1. Mean Squared Error(*MSE*) :

$$MSE = \sum_{j=1}^N (observation_j - prediction_j)^2 / N$$

2. Root Mean Squared Error(*RMSE*) :

$$RMSE = \sqrt{\sum_{j=1}^N (observation_j - prediction_j)^2 / N}$$

3. Normalized Mean Squared Error(*NMSE*) :

$$\sum_{j=1}^N (observation_j - prediction_j)^2 / \sum_{j=1}^N (observation_j - mean)^2$$

4. Mean Absolute Percentage Error(*MAPE*) :

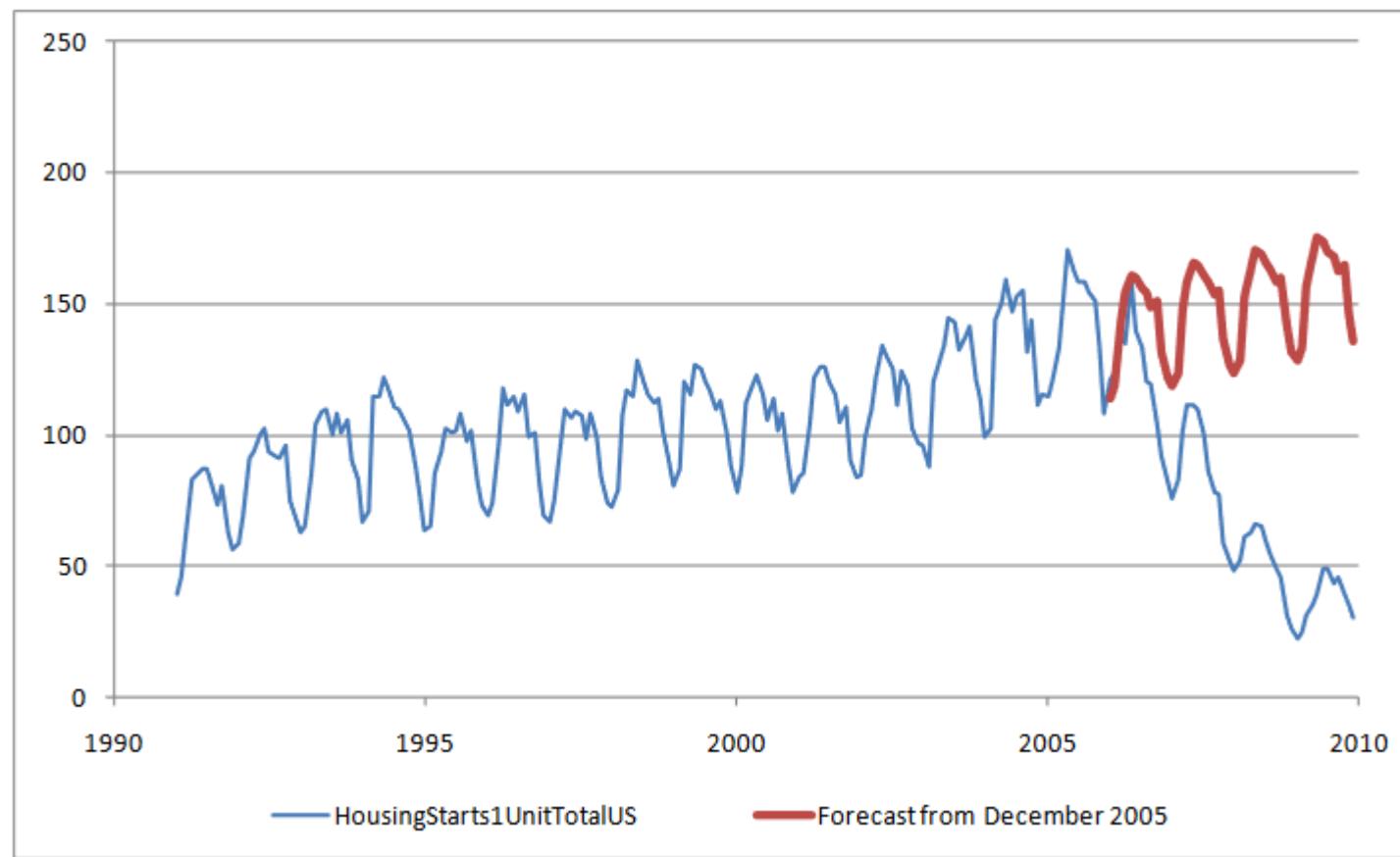
$$MAPE = 100 \times \frac{\sum_{i=0}^N (|forecasted_i - actual_i|) / actual_i}{N}$$

Rodzaje ryzyka predykcji

- Ryzyko wewnętrzne (intrinsic) – zaszumienie danych jest zbyt duże żeby znaleźć wzorce
 - występuje zawsze
 - można oszacować (“standard error”)
 - redukcja przez poszukiwanie wzorców → trendów, sezonowości itp. (więcej danych pomaga)
- Ryzyko doboru parametrów – możliwości analiz są olbrzymie
 - mierzony przez “standard errors of the coefficient estimates”
 - redukcja przez zastosowanie lepiej dopasowanych metod
 - więcej danych może nie pomóc → *No pattern really stays the same forever*
- Ryzyko wyboru modelu – przyjęcie błędnego założenia o prawdziwości modelu (np. do predykcji)
 - redukcja przez analizę danych, sprawdzenie założeń itp.
 - żadna analiza nie dostarczy oszacowania takiego błędu
 - mierzalny skutkami

Modele predykcyjne - ryzyko

“If you live by the crystal ball you end up eating broken glass”





INNE CELE ANALIZ

Inne cele analizy szeregu czasowego

Porównywanie szeregów

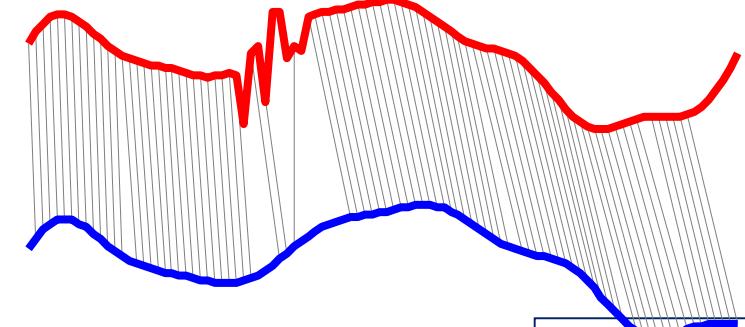
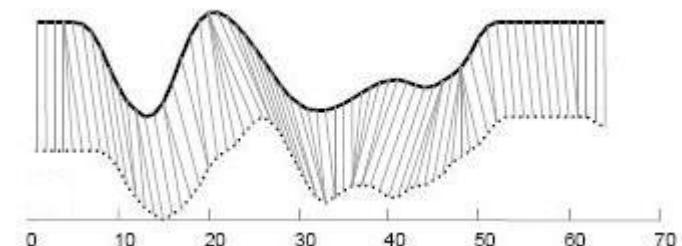
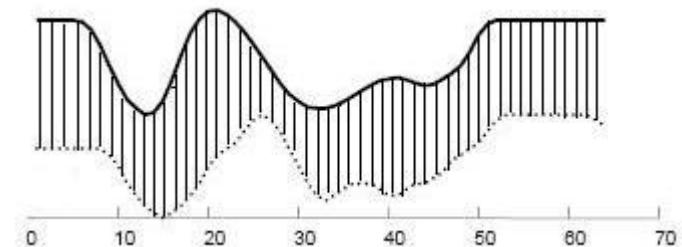
Wykrywanie motywów

Wykrywanie anomalii

- Dodatkowe możliwości przekształceń
 - Stacjonaryzacja szeregu
 - Zmiana reprezentacji szeregu
 - Analiza w dziedzinie częstotliwości
 - Analiza w dziedzinie czasu-częstotliwości

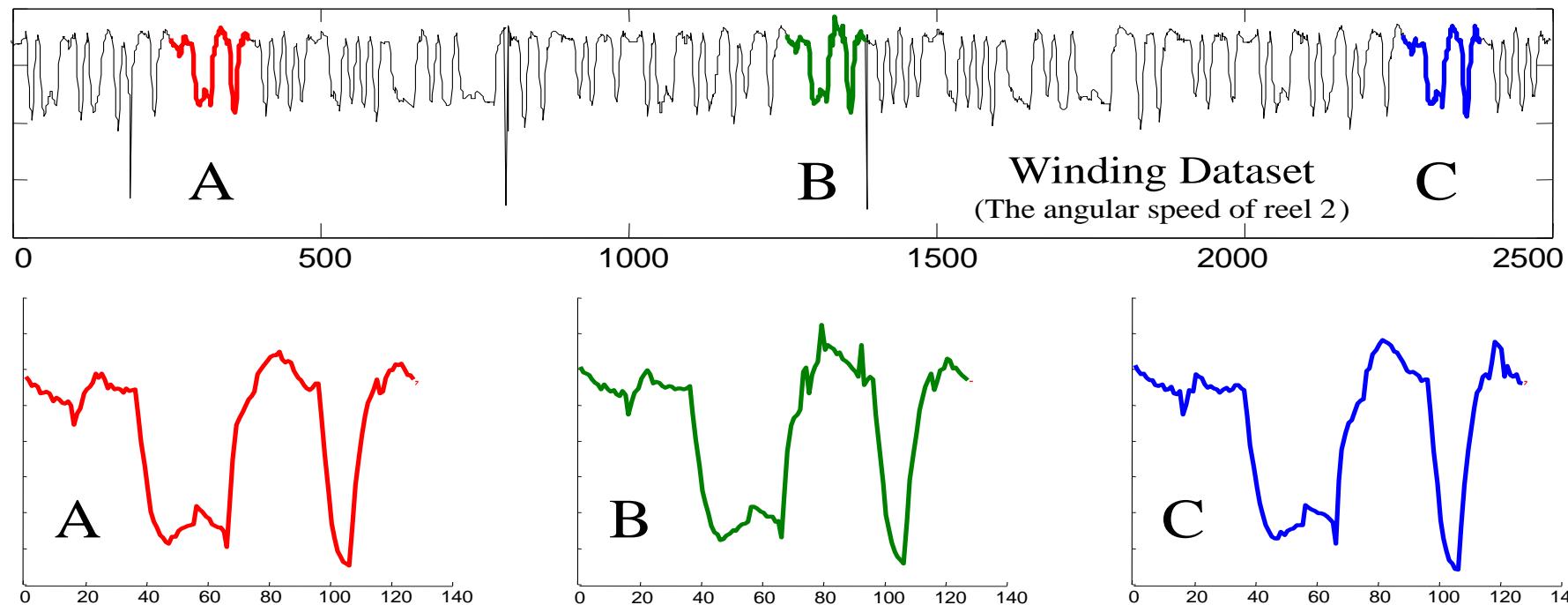
Porównywanie szeregów

- Odległość euklidesowa
- DTW (Dynamic Time Warping)
- LCSS (Longest Common Subsequence)

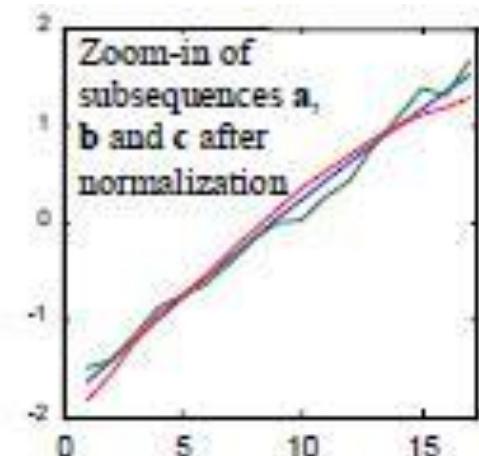
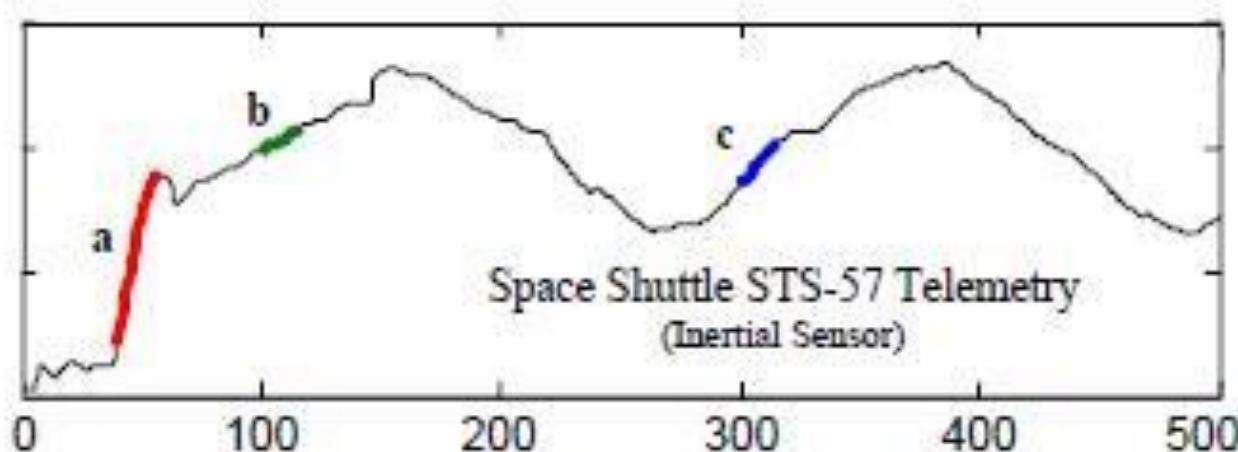


Wykrywanie motywów

Motyw – powtarzalny schemat przebiegu szeregu czasowego

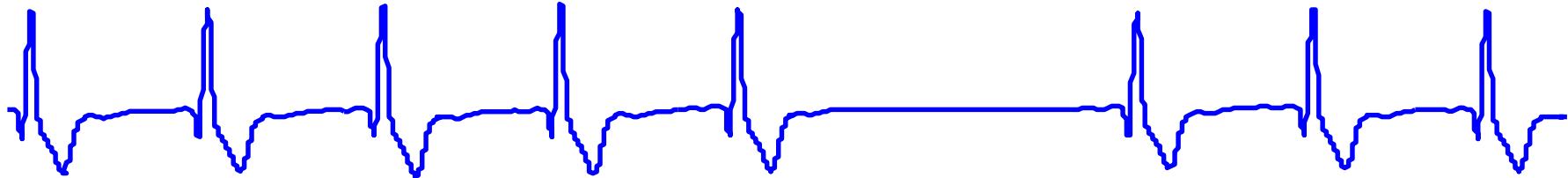


Wykrywanie motywów *problemy*



Wykrywanie anomalii

Anomalia – fragment szeregu czasowego odbiegający od jego typowego przebiegu



Aberrant Behavior

Novelties

Anomalies

Faults

Surprises

Deviants

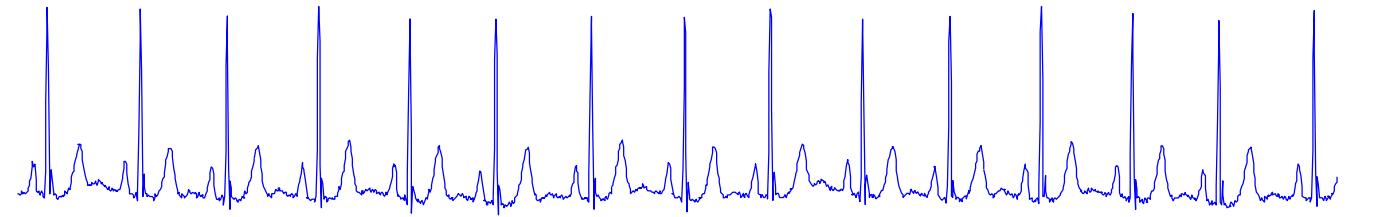
Temporal Change

Outliers

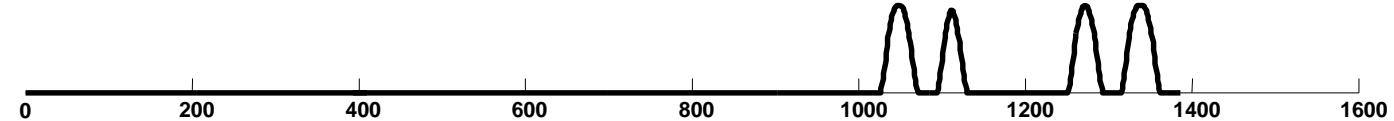
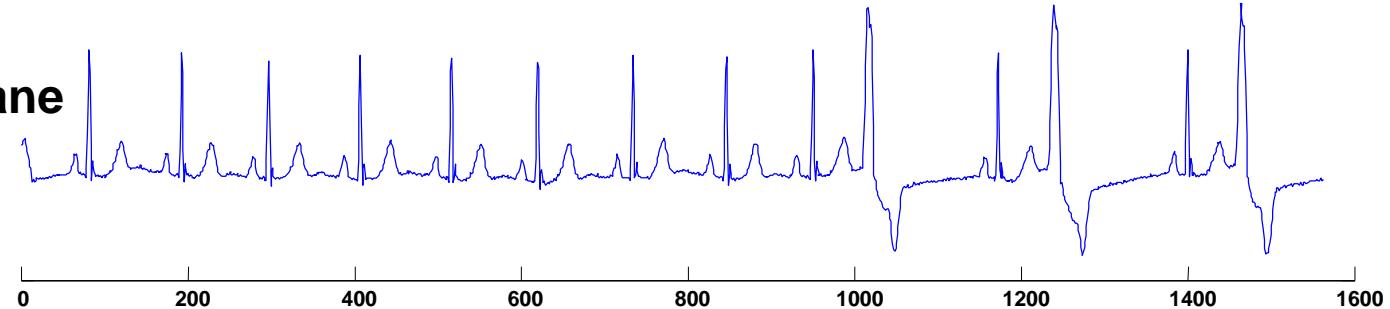
Fragment można uznać za nietypowy jeżeli jego **częstość występowania różni się** znacząco od częstości, której spodziewamy się na podstawie danych.

Wykrywanie anomalii

Dane uczące

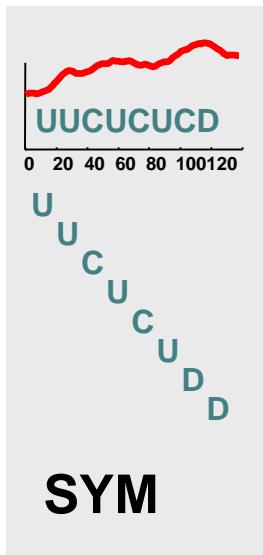
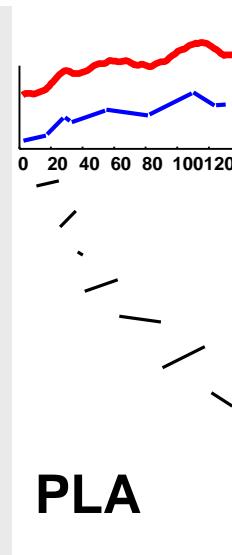
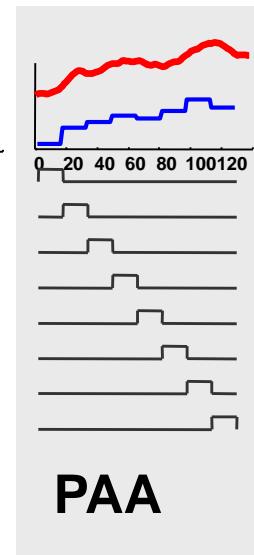
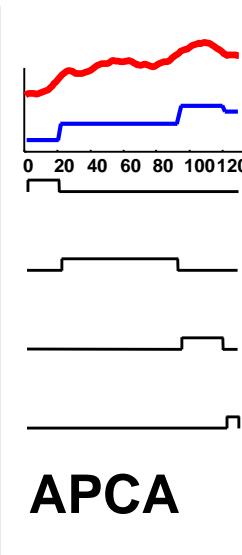
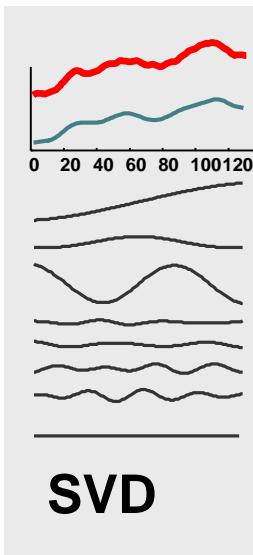
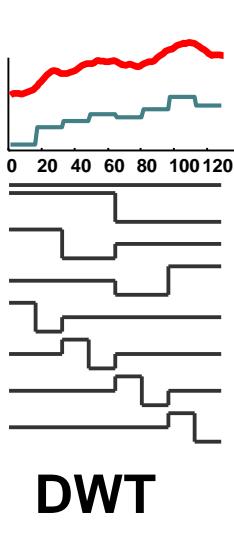
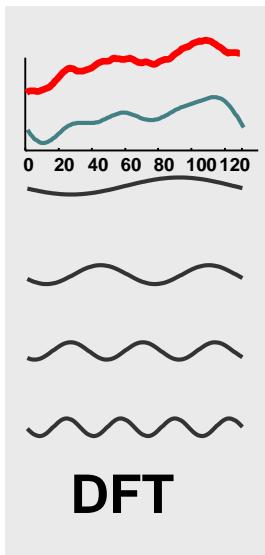


Dane analizowane



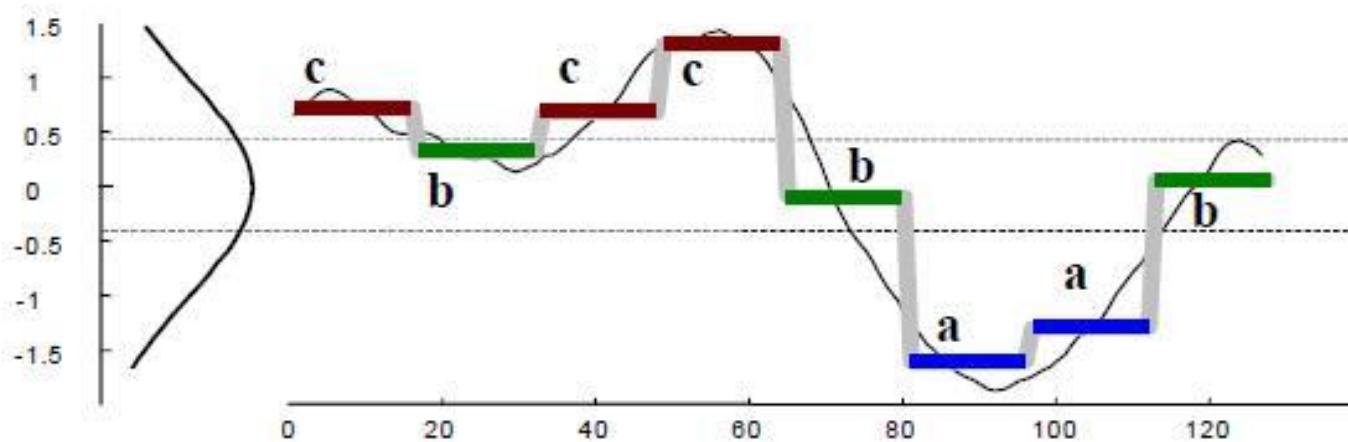
Alternatywne reprezentacje szeregów czasowych

- Discrete Fourier Transformation
- Discrete Wavelet Transformation
- Singular Value Decomposition
- Adaptive Piecewise Constant Approximation
- Piecewise Aggregate Approximation
- Piecewise Linear Approximation
- Symboliczna (np. SAX)



Symbolic Aggregate Approximation SAX

1. Redukcja wymiaru (Piecewise Aggregate Approximation)
2. Dyskretyzacja (przypisanie symboli)



cbccbaab



Szeregi czasowe - podsumowanie

- Znany cel analizy
- Ocena szeregu **ZANIM** zabierzemy się za analizę i predykcję
 - Źródło, częstotliwość, jednostka itp.
 - Kompletność, wartości odstające
- Analiza wstępna surowych danych
- Przekształcenia
- Budowa modeli predykcyjnych/Szukanie wzorców ...
- Zdefiniowane metryki oceny i kryteria akceptacji
- Ocena i oszacowanie ryzyka prawdziwości wyników
- Języki do analizy: R, Python
- Dalsze kroki:
 - Analizy w dziedzinie częstotliwości oraz czasu-częstotliwości
 - Okienkowanie i inne przekształcenia szeregów
 - Obliczanie wektorów cech i uczenie maszynowe (rozpoznawanie, detekcja)