Wyrażenia regularne

dr inż. Marcin Ciura
marcin.ciura@uken.krakow.pl

Wydział Bezpieczeństwa i Informatyki UKEN

Plan na dziś

- Wyrażenia regularne
- Automaty
- Regexpy w Pythonie
- ...a ponadto wiadomości o 4 ludziach i kilka zagadek

Wyrażenia regularne

Do czego służą wyrażenia regularne?

Wyrażenia regularne służą do wyszukiwania wzorców w tekstach

Przykład

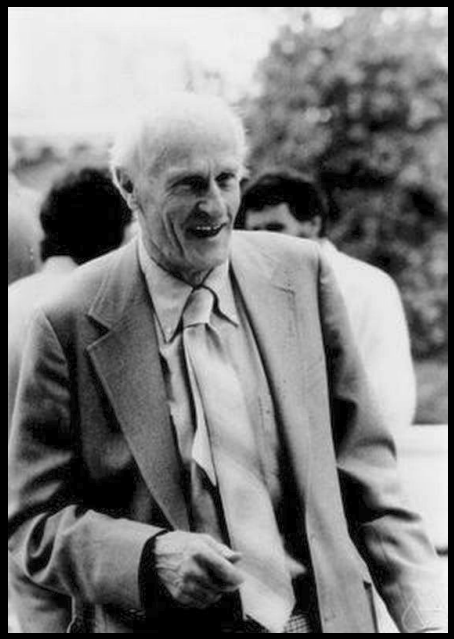
Każdy gen RNA pasuje do tego wyrażenia regularnego:

`(AUG|CUG|UUG)(...)*(UAA|UAG|UGA)`

Wyrażenia regularne a regex(p)y

- Wyrażenia regularne to pojęcie matematyczne
- Regexpy, znane też jako regexy, to implementacja wyrażeń regularnych w pewnym języku programowania wraz z dodatkami, które ułatwiają pracę programistom

Stephen Cole Kleene (5.1.1909–25.1.1994)



Amerykański matematyk, znany z pewnej gwiazdki. Stworzył teorię obliczalności. Wspinał się po górach, działał na rzecz ochrony przyrody. W 1951 roku, kiedy badał sieci neuronowe, wynalazł wyrażenia regularne

Definicje

- Alfabet: skończony zbiór symboli czyli znaków,
  na przykład {A, C, G, T}, ASCII, Unicode...
- Łańcuch znaków, krócej łańcuch: to samo, co ciąg znaków,
  na przykład ATGTGA
- Długość łańcucha: liczba znaków w tym łańcuchu. Oznaczam długość łańcucha, obejmując ten łańcuch kreskami pionowymi,
  na przykład $|\texttt{ATGTGA}| = 6$
- Łańcuch pusty: łańcuch o długości 0 znaków. Oznaczam łańcuch pusty grecką literą epsilon: ε

Definicje

- Wyrażenie regularne to taki łańcuch, który opisuje pewien zbiór łańcuchów zgodnie z pewnymi regułami Mówimy, że zbiór łańcuchów opisany przez pewne wyrażenie regularne pasuje do tego wyrażenia regularnego
  Możemy też powiedzieć o każdym łańcuchu z tego zbioru, że ten łańcuch pasuje do danego wyrażenia regularnego

Elementarne wyrażenia regularne

Każde elementarne wyrażenie regularne składa się z jednego symbolu

- Do symbolu zbioru pustego Ø pasuje pusty zbiór łańcuchów: Ø
- Do symbolu pustego łańcucha ε pasuje zbiór {ε}, który zawiera tylko łańcuch pusty
- Do każdego wyrażenia regularnego złożonego z jednego znaku pasuje jeden łańcuch o długości 1. Ten łańcuch składa się z tego samego znaku, co wyrażenie regularne
  na przykład do wyrażenia regularnego a pasuje zbiór łańcuchów {a} :-)

Złożone wyrażenia regularne

- nawiasy
- konkatenacja wyrażeń regularnych
- alternatywa wyrażeń regularnych
- gwiazdka Kleene'a

Złożone wyrażenia regularne: nawiasy

Jeśli *R* jest wyrażeniem regularnym, to:

- wyrażenie regularne *(R)* oznacza to samo, co wyrażenie regularne *R* :-)

Złożone wyrażenia regularne: konkatenacja

Jeśli R i S są wyrażeniami regularnymi, to:

- Do wyrażenia regularnego RS, czyli do konkatenacji wyrażeń regularnych R i S, pasują takie łańcuchy, które powstają, gdy łączę dowolny łańcuch pasujący do R z dowolnym łańcuchem pasującym do S

Wyraz konkatenacja pochodzi od angielskiego czasownika *to concatenate*, czyli łączyć w łańcuch

Złożone wyrażenia regularne: konkatenacja

Przykłady konkatenacji wyrażeń regularnych:

- Jeśli do R pasuje zbiór łańcuchów {d},
  a do S pasuje zbiór łańcuchów {o},
  to do RS pasuje zbiór łańcuchów {do}
- Jeśli do R pasuje zbiór łańcuchów {do, od},
  a do S pasuje zbiór łańcuchów {dać, pisać},
  to do RS pasuje zbiór łańcuchów {dodać, dopisać, oddać, odpisać}

Złożone wyrażenia regularne: alternatywa

Jeśli R i S są wyrażeniami regularnymi, to:

- Do wyrażenia regularnego $R|S$, czyli do alternatywy wyrażeń regularnych R i S, pasuje suma dwóch zbiorów: zbioru takich łańcuchów, które pasują do R i zbioru takich łańcuchów, które pasują do S

Złożone wyrażenia regularne: alternatywa

Przykłady alternatywy wyrażeń regularnych:

- Jeśli do R pasuje zbiór łańcuchów $\{\texttt{do}\}$,
  a do S pasuje zbiór łańcuchów $\{\texttt{od}\}$,
  to do $R|S$ pasuje zbiór łańcuchów $\{\texttt{do}, \texttt{od}\}$
- Jeśli do R pasuje zbiór łańcuchów $\{\texttt{ręka}, \texttt{noga}\}$,
  a do S pasuje zbiór łańcuchów $\{\texttt{ręka}, \texttt{głowa}\}$,
  to do $R|S$ pasuje zbiór łańcuchów $\{\texttt{ręka}, \texttt{noga}, \texttt{głowa}\}$

Złożone wyrażenia regularne: gwiazdka Kleene'a

Jeśli R jest wyrażeniem regularnym, to:

- Do wyrażenia regularnego R^*, czyli do domknięcia Kleene'a wyrażenia regularnego R pasuje zbiór takich łańcuchów, które powstają, gdy łączę 0 lub więcej łańcuchów pasujących do R
  Domknięcie Kleene'a nazywa się również gwiazdką Kleene'a

Złożone wyrażenia regularne: gwiazdka Kleene'a

Przykłady użycia gwiazdki Kleene'a:

- Jeśli do R pasuje zbiór łańcuchów `{A}`, to do R^* pasuje zbiór łańcuchów
  `{ε, A, AA, AAA, AAAA, AAAAA, AAAAAA, AAAAAAA, ...}`
- Jeśli do R pasuje zbiór łańcuchów `{fa, sol}`, to do R^* pasuje zbiór łańcuchów

```
{ε,
    fa, sol,
    fafa, fasol, solfa, solsol,
    fafafa, fafasol, fasolfa, fasolsol,
    solfafa, solfasol, solsolfa, solsolsol,
    ...
}
```

Złożone wyrażenia regularne: kolejność działań

Wykonuję działania w takiej kolejności:

- najpierw wykonuję działania w nawiasach
- potem stosuję gwiazdkę Kleene'a
- potem konkatenuję wyrażenia regularne
- potem buduję alternatywy wyrażeń regularnych

Zagadka

Które z dwóch wyrażeń regularnych po prawej stronie znaku = jest równoważne wyrażeniu regularnemu po lewej stronie znaku =?

1. `ab*` = `a(b*)` czy `(ab)*`?
2. `a|b*` = `a|(b*)` czy `(a|b)*`?
3. `ab|cd` = `(ab)|(cd)` czy `a(b|c)d`?

Kolejność działań: nawiasy, gwiazdka, konkatenacja, alternatywa

Zagadka: rozwiązanie

Które z dwóch wyrażeń regularnych po prawej stronie znaku = jest równoważne wyrażeniu regularnemu po lewej stronie znaku =?

1. `ab*` = `a(b*)`
2. `a|b*` = `a|(b*)`
3. `ab|cd` = `(ab)|(cd)`

Kolejność działań: nawiasy, gwiazdka, konkatenacja, alternatywa

Zagadka

Które łańcuchy pasują do wyrażenia regularnego `(b*(a|ε)b)*`?

1. ε
2. a
3. b
4. c
5. ab
6. aa
7. bb
8. ba
9. bbbbb
10. bbbba
11. abbbb
12. aaabb
13. bbabb
14. baabb
15. ababab

Zagadka: rozwiązanie

Które łańcuchy pasują do wyrażenia regularnego `(b*(a|ε)b)*`?

1. ε
2. a
3. b
4. c
5. ab
6. aa
7. bb
8. ba
9. bbbbb
10. bbbba
11. abbbb
12. aaabb
13. bbabb
14. baabb
15. ababab

Proszę Państwa o pytania :-)

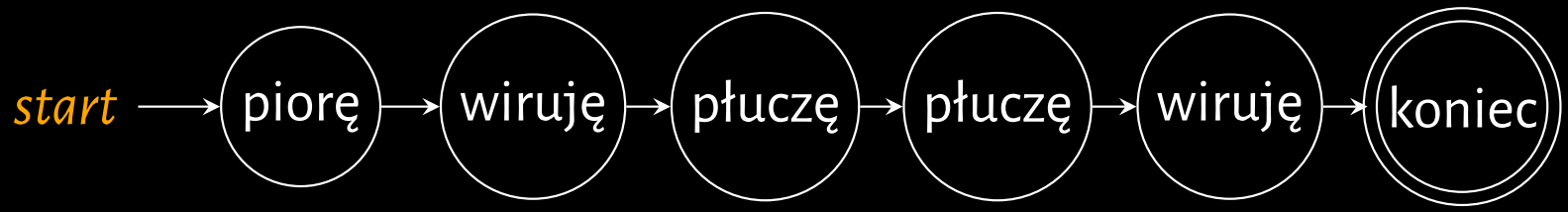
Automaty

Edward Forrest Moore (23.11.1925–14.6.2003)

Amerykański matematyk i informatyk. W 1956 roku zaproponował konstrukcję takich maszyn, które same by się rozmnażały. Przypuszczał, że takie maszyny można by produkować łatwiej niż statki kosmiczne. W tym samym roku wymyślił taki rodzaj automatu, o jakim będzie mowa na wykładzie

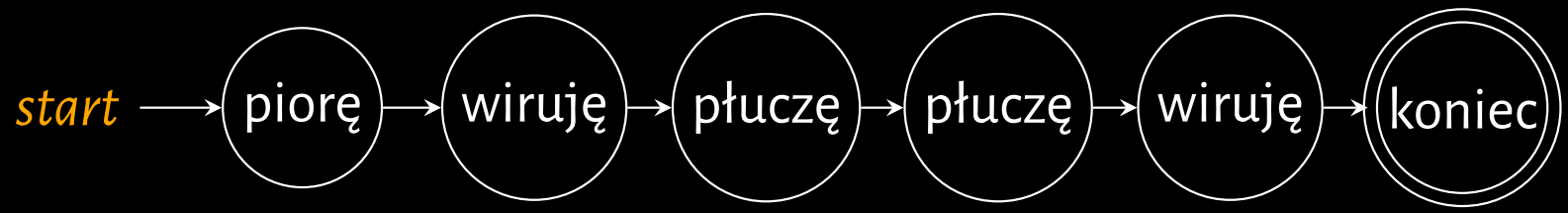
Jak działa automat?

Jak działa automat?



Automat wie, co robi, ale nie pamięta, jak do tego doszedł

Jak działa automat?



Kółka oznaczają stany automatu

Strzałki oznaczają przejścia między stanami automatu :-)

Automat wie, w jakim stanie jest, ale nie pamięta, jak do niego doszedł

Automaty

Automaty w informatyce różnią się od pralki automatycznej tym, że wczytują znaki. Każdy automat zmienia swój stan zgodnie z tym znakiem, który wczytał.

Każde wyrażenie regularne można przekształcić na automat

Automaty skończone

Dalej będzie mowa o automatach skończonych, czyli takich automatach, które mają skończoną liczbę stanów

Niedeterministyczne automaty skończone

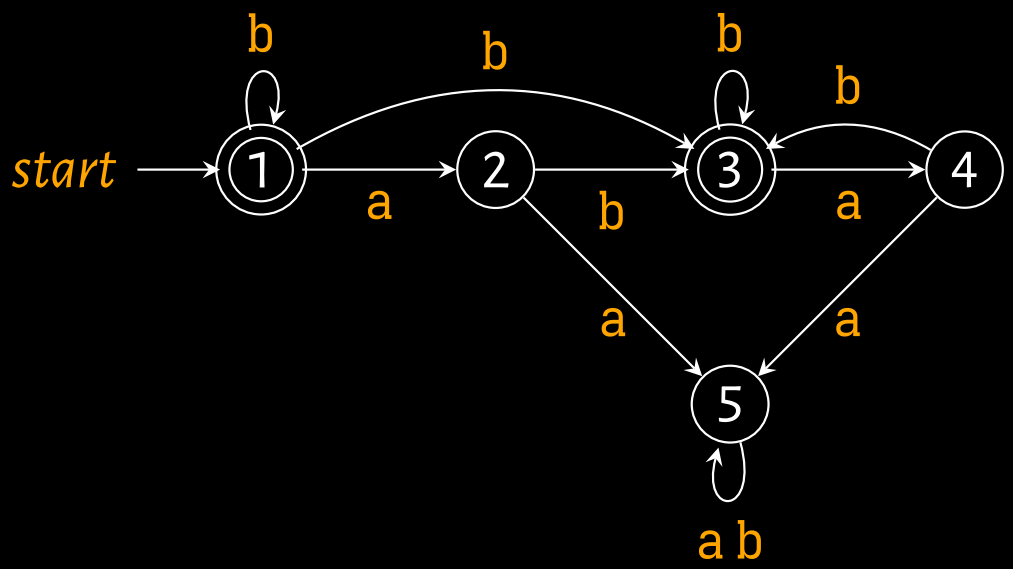
NFA (niedeterministyczny automat skończony, po angielsku: Nondeterministic Finite Automaton)

Z jednego stanu NFA może wychodzić więcej niż jedno przejście oznaczone tak samo

Gdy NFA znajduje się w pewnym stanie i wczytuje pewien znak, może przejść do jednego z wielu różnych stanów

Przykład NFA

(b*(a|ε)b)*



Deterministyczne automaty skończone

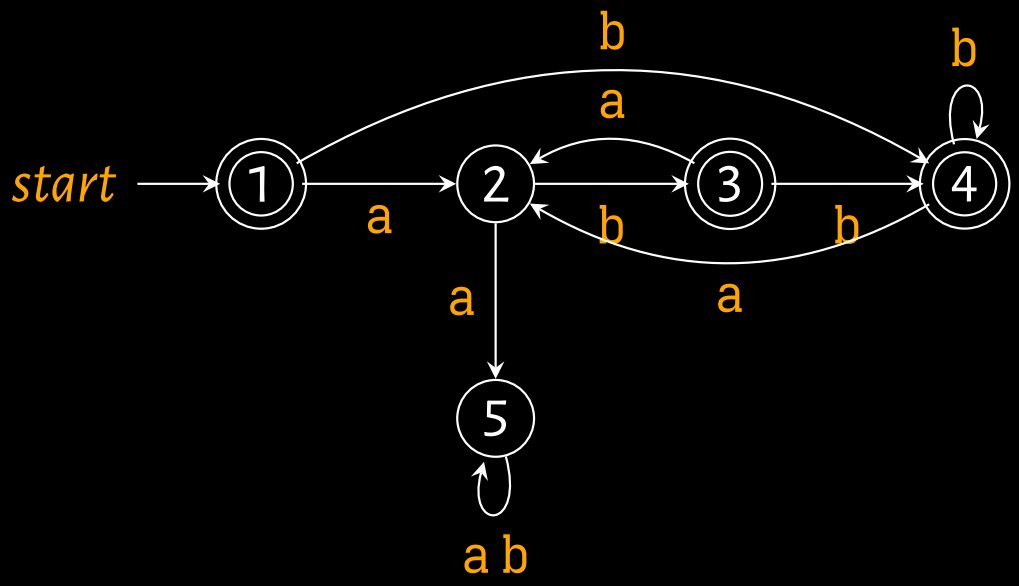
DFA (deterministyczny automat skończony, po angielsku: Deterministic Finite Automaton)

Z jednego wierzchołka grafu, który odpowiada danemu DFA, wychodzi dokładnie jedna krawędź oznaczona danym znakiem

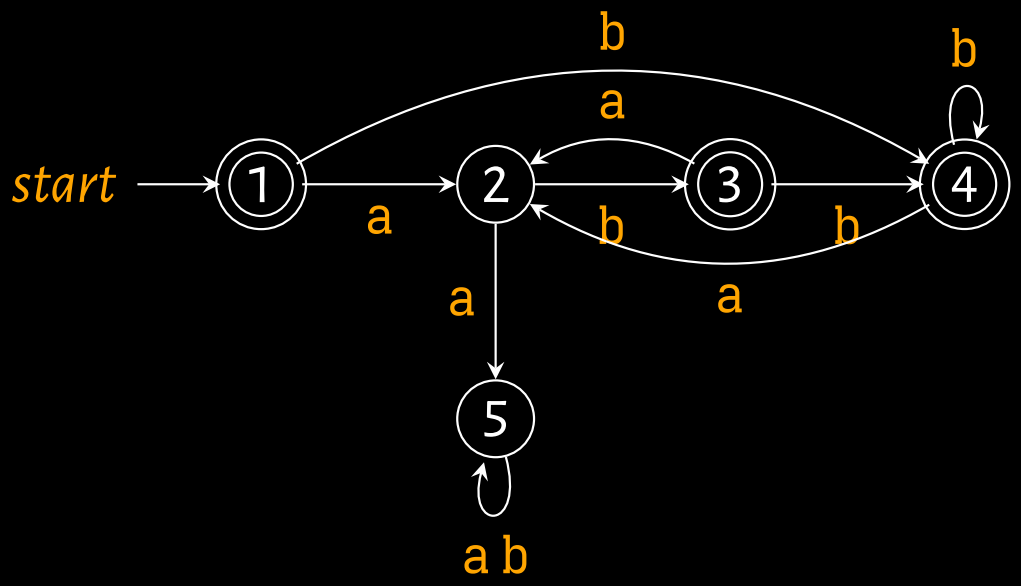
Gdy DFA znajduje się w pewnym stanie i wczytuje pewien znak, zawsze przechodzi do określonego stanu

Przykład DFA

$(b^*(a|\varepsilon)b)^*$



DFA: tabela przejść i wyjść

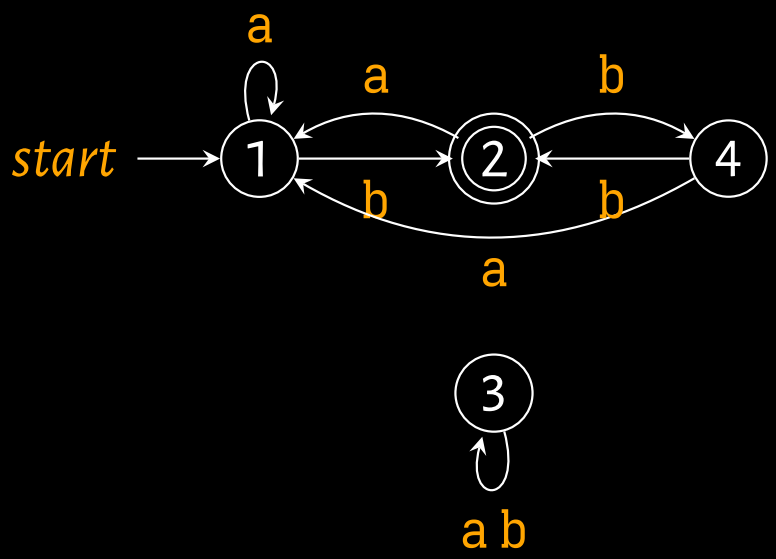


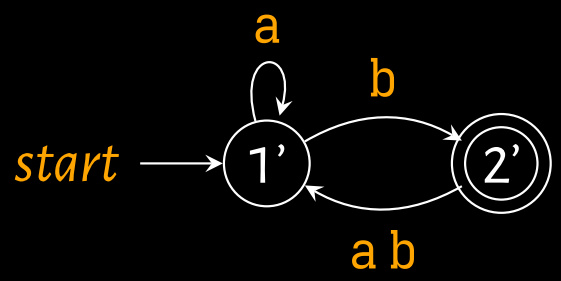
stan	końcowy?	a	b
1	true	2	4
2	false	3	5
3	true	2	4
4	true	2	4
5	false	5	5

Równoważne automaty skończone

Mówimy, że 2 automaty skończone są równoważne, jeśli rozpoznają te same zbiory łańcuchów

Przykład 2 równoważnych automatów skończonych, NFA i DFA:





Jak dopasować łańcuch do wyrażenia regularnego?

Aby dopasować łańcuch do danego wyrażenia regularnego, można:

- Bezpośrednio interpretować to wyrażenie regularne
- Zbudować NFA, który odpowiada temu wyrażeniu regularnemu (na przykład algorytmem Thompsona), po czym:
  - przechodzić przez ten NFA, pamiętając bieżący zbiór stanów
  - przechodzić przez ten NFA z nawrotami (po angielsku:backtracking)
  - zbudować DFA równoważny temu NFA
- Od razu zbudować DFA, który odpowiada temu wyrażeniu regularnemu, korzystając z:
  - algorytmu Myhilla-Nerode'a
  - algorytmu DeRemera
  - pochodnych Brzozowskiego

Janusz Brzozowski (10.5.1935–24.10.2019)

Polsko-kanadyjski informatyk. Urodził się w Warszawie. W 1964 roku wymyślił pochodne Brzozowskiego. Pochodne Brzozowskiego są eleganckie, bo opierają się na prostym pomyśle

Proszę Państwa o pytania :-)

Regexpy w Pythonie

Regexpy

Regexpy, znane też jako regexy, to implementacja wyrażeń regularnych w danym języku programowania wraz z dodatkami, które ułatwiają pracę programistom

Tę część wykładu opracowałem na podstawie dokumentacji modułu `re`

Dokumentacja tego pakietu znajduje się pod adresem `https://docs.python.org/3/library/re.html` i na dysku lokalnym

Widzę tę dokumentację na ekranie, gdy w interpreterze Pythona wydaję polecenia

```
import re
help(re)
```

Dziękuję Państwu za uwagę :-)

Źródła zdjęć

https://en.wikipedia.org/wiki/Stephen_Cole_Kleene

https://my-concept.pl/pl/p/Pralka-automatyczna-SLIM-PP6306s/20820

https:
//en.wikipedia.org/wiki/Janusz_Brzozowski_(computer_scientist)

https://www.azquotes.com/author/16151-Jamie_Zawinski