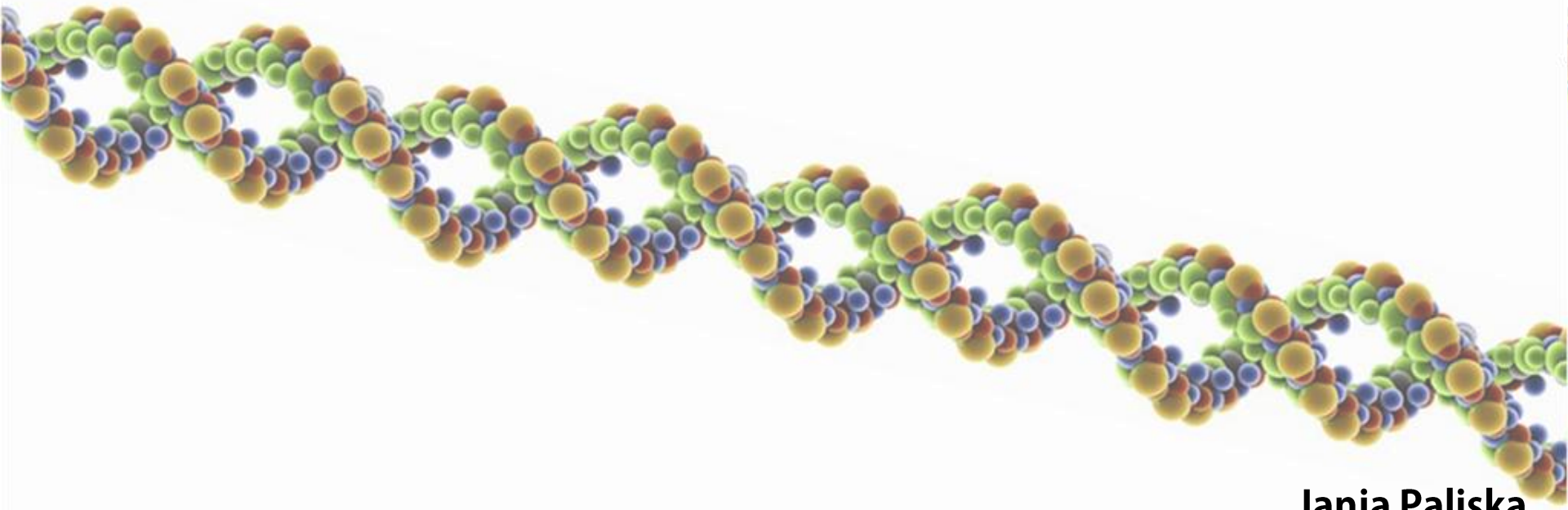
PORAVNAVANJE KRATKIH NIZOVA S DUGAČKIM KORISTEĆI DFA (AHO-CORASICK ALGORITAM)



Janja Paliska
Antonio Soldo
Antonio Zemunik

Uvod

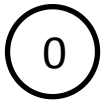
- K = $\{y_1, y_2, ..., y_k\}$ → skup ključnih riječi
- x → tekst u kojem tražimo ključne riječi
- Kako efikasno pronaći sve ključne riječi jednim prolaskom kroz tekst?
  - koristeći DFA i Aho-Corasick algoritam

Aho-Corasick algoritam

- ponašanje DFA određeno sljedećim funkcijama:
  - *goto*: g(s, a) = s' / fail
  - *failure*: f(s) = s_f,
  - *output*: output(s) = {y_1, y_2, ..., y_n}

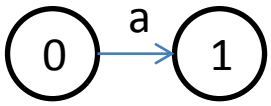
Primjer rada Aho-Corasick algoritma

K = {abc, bc, ac}



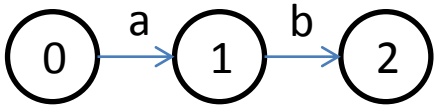
Primjer rada Aho-Corasick algoritma

K = {abc, bc, ac}



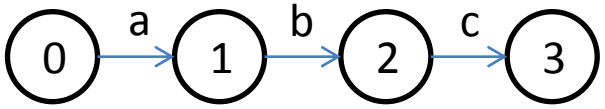
Primjer rada Aho-Corasick algoritma

K = {abc, bc, ac}



Primjer rada Aho-Corasick algoritma

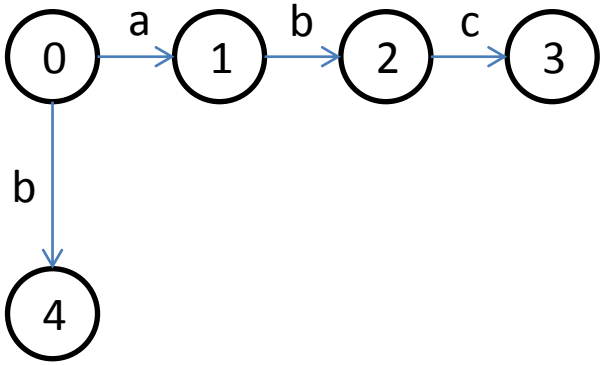
K = {abc, bc, ac}



s	output(s)
3	abc

Primjer rada Aho-Corasick algoritma

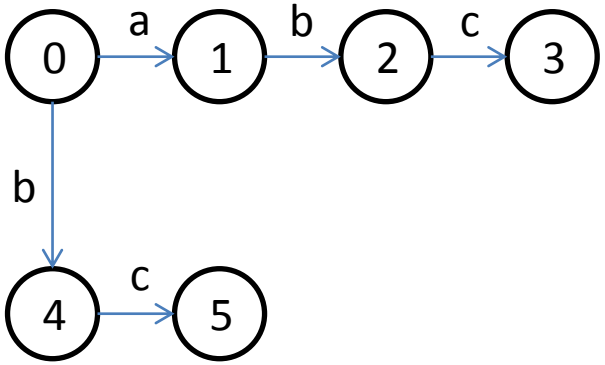
K = {abc, bc, ac}



s	output(s)
3	abc

Primjer rada Aho-Corasick algoritma

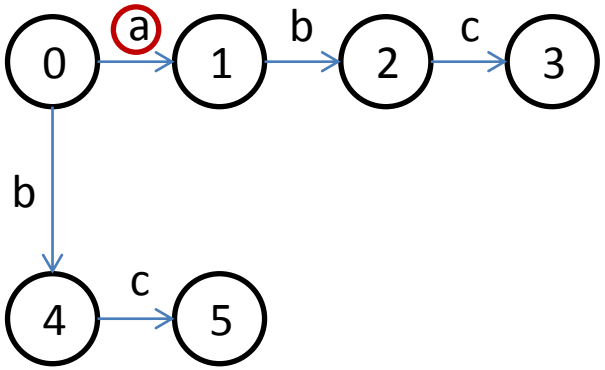
K = {abc, bc, ac}



s	output(s)
3	abc
5	bc

Primjer rada Aho-Corasick algoritma

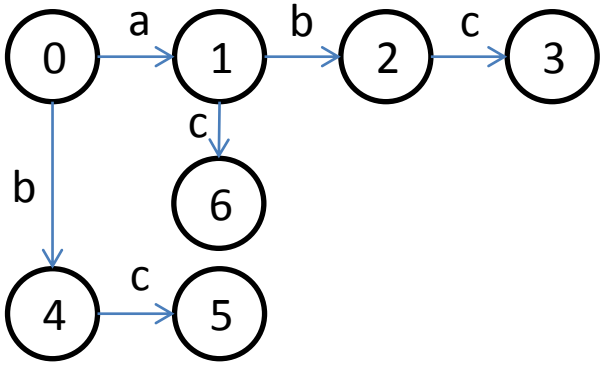
K = {abc, bc, ac}



s	output(s)
3	abc
5	bc

Primjer rada Aho-Corasick algoritma

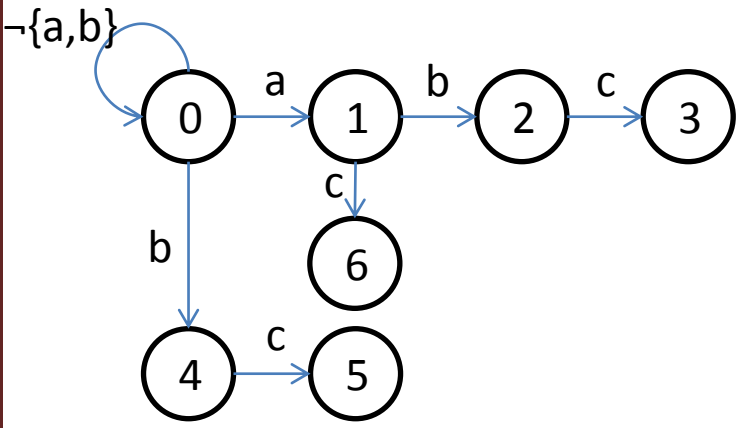
K = {abc, bc, ac}



s	output(s)
3	abc
5	bc
6	ac

Primjer rada Aho-Corasick algoritma

K = {abc, bc, ac}

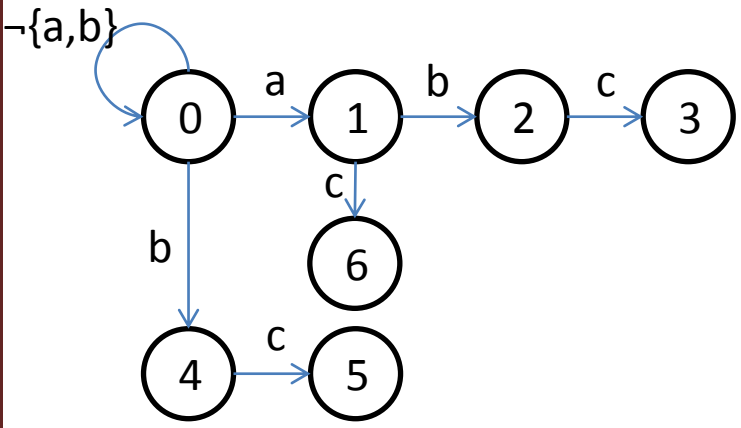


s	output(s)
3	abc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

K = {abc, bc, ac}

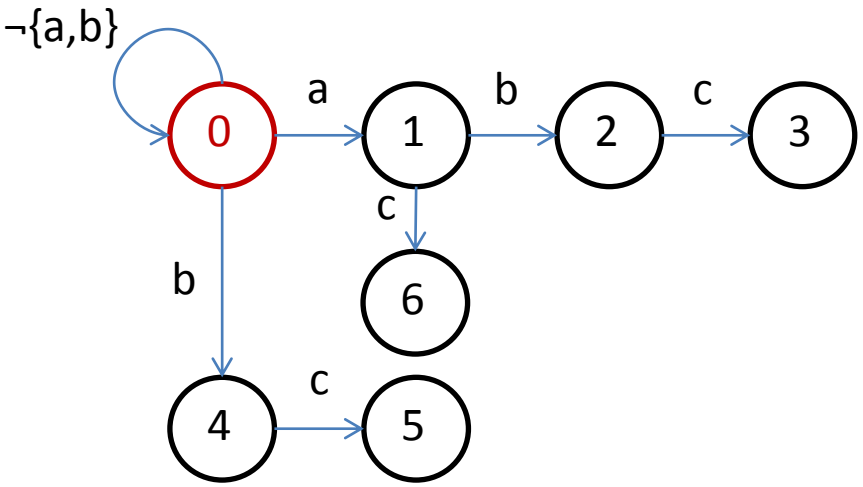


s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia

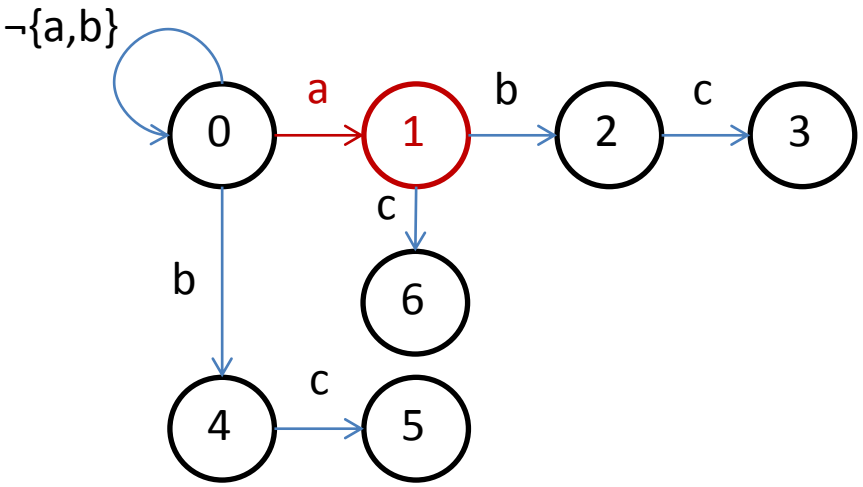


s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia

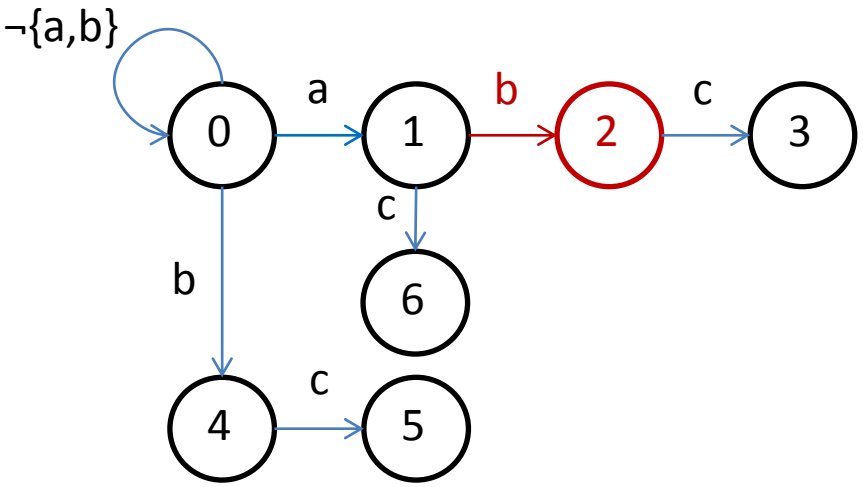


s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia

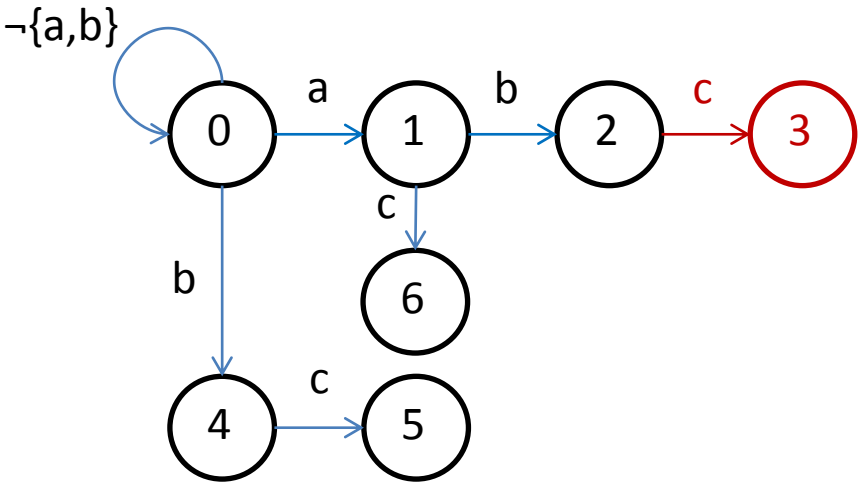


s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia



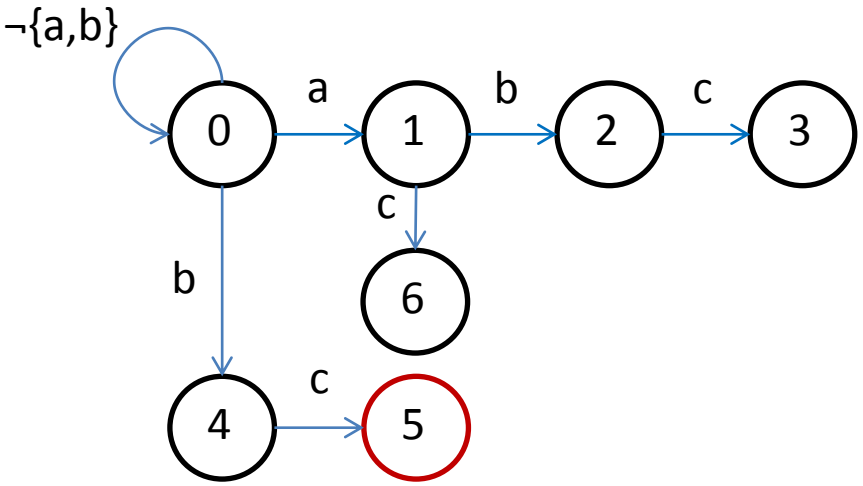
abc, bc

s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia



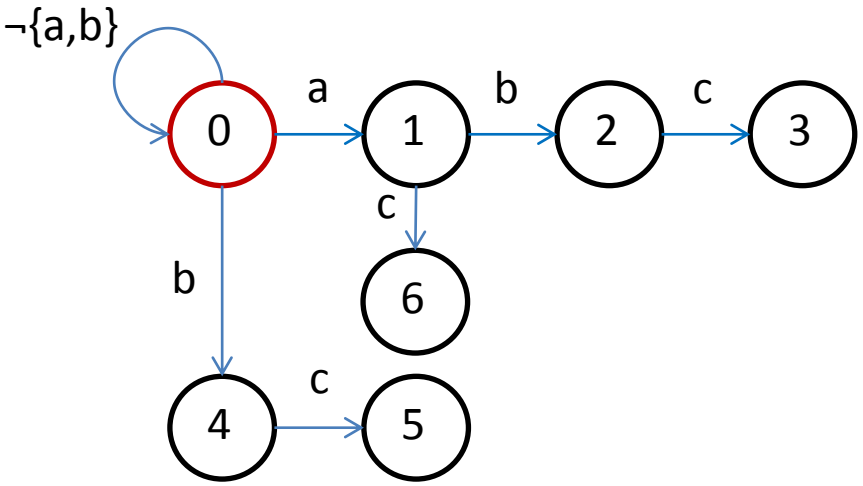
abc, bc

s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia



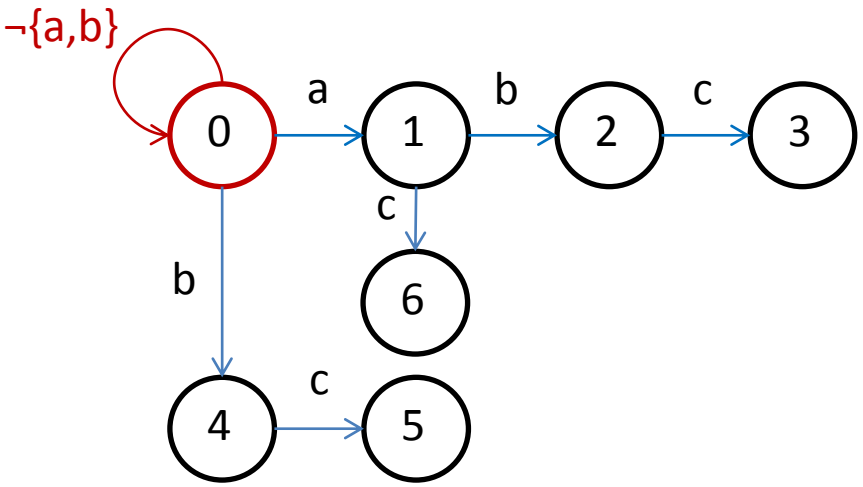
abc, bc

s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia



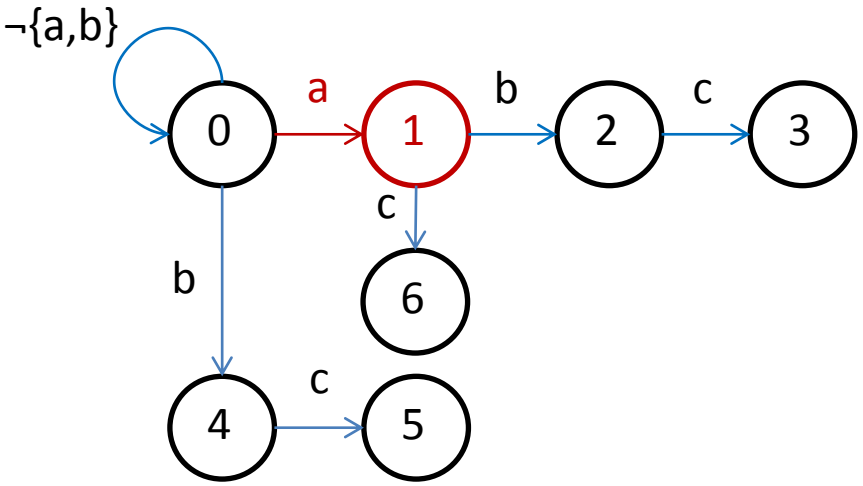
abc, bc

s	output(s)
3	abc, bc
5	bc
6	ac

s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Primjer rada Aho-Corasick algoritma

text = abcia



abc, bc

s	output(s)
3	abc, bc
5	bc
6	ac

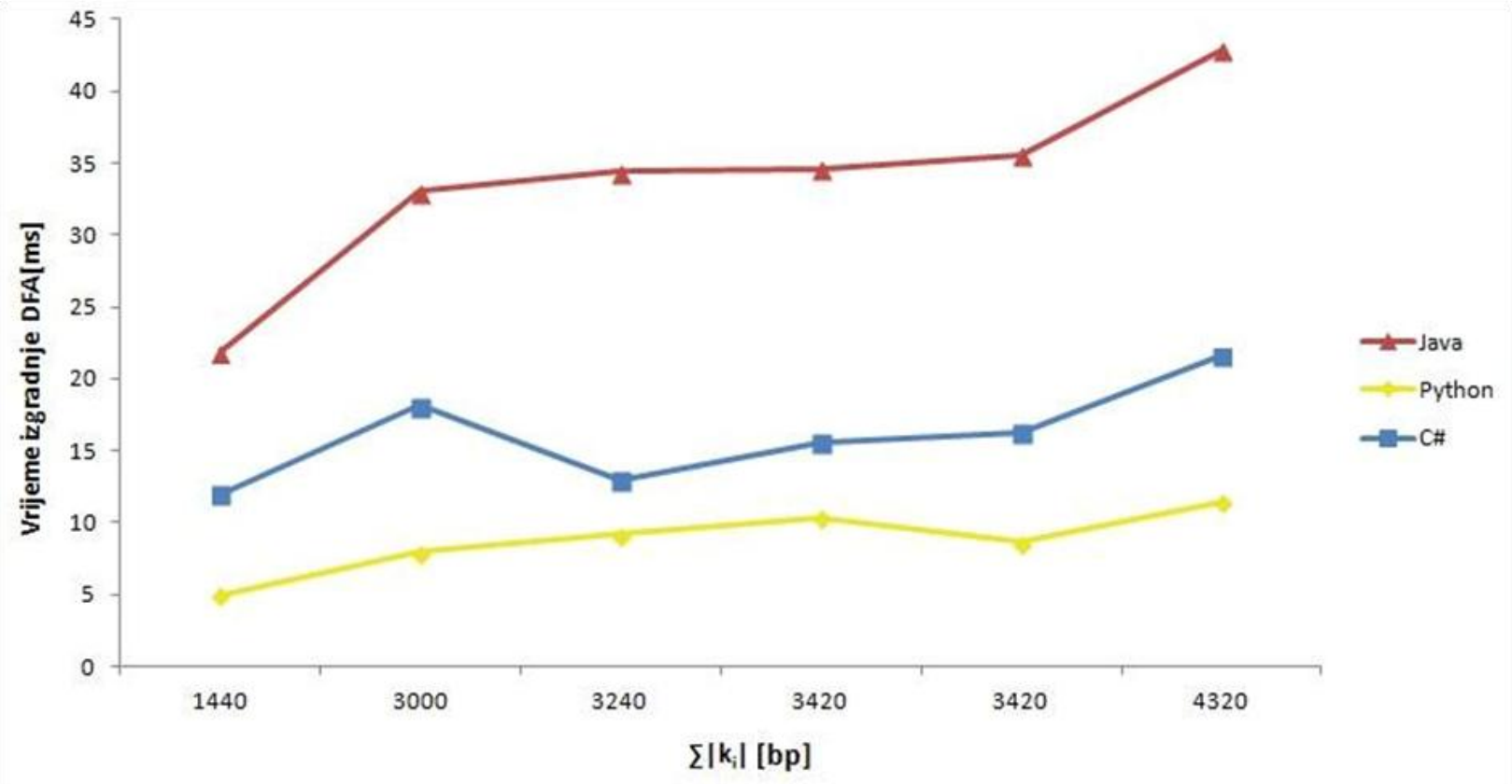
s	f(s)
1	0
2	4
3	5
4	0
5	0
6	0

Implementacija i rezultati

- implementacija: Java, Python, C#
- mjereni parametri:
  a) vrijeme izgradnje DFA
  b) vrijeme prolaska teksta kroz DFA
  c) utrošena memorija

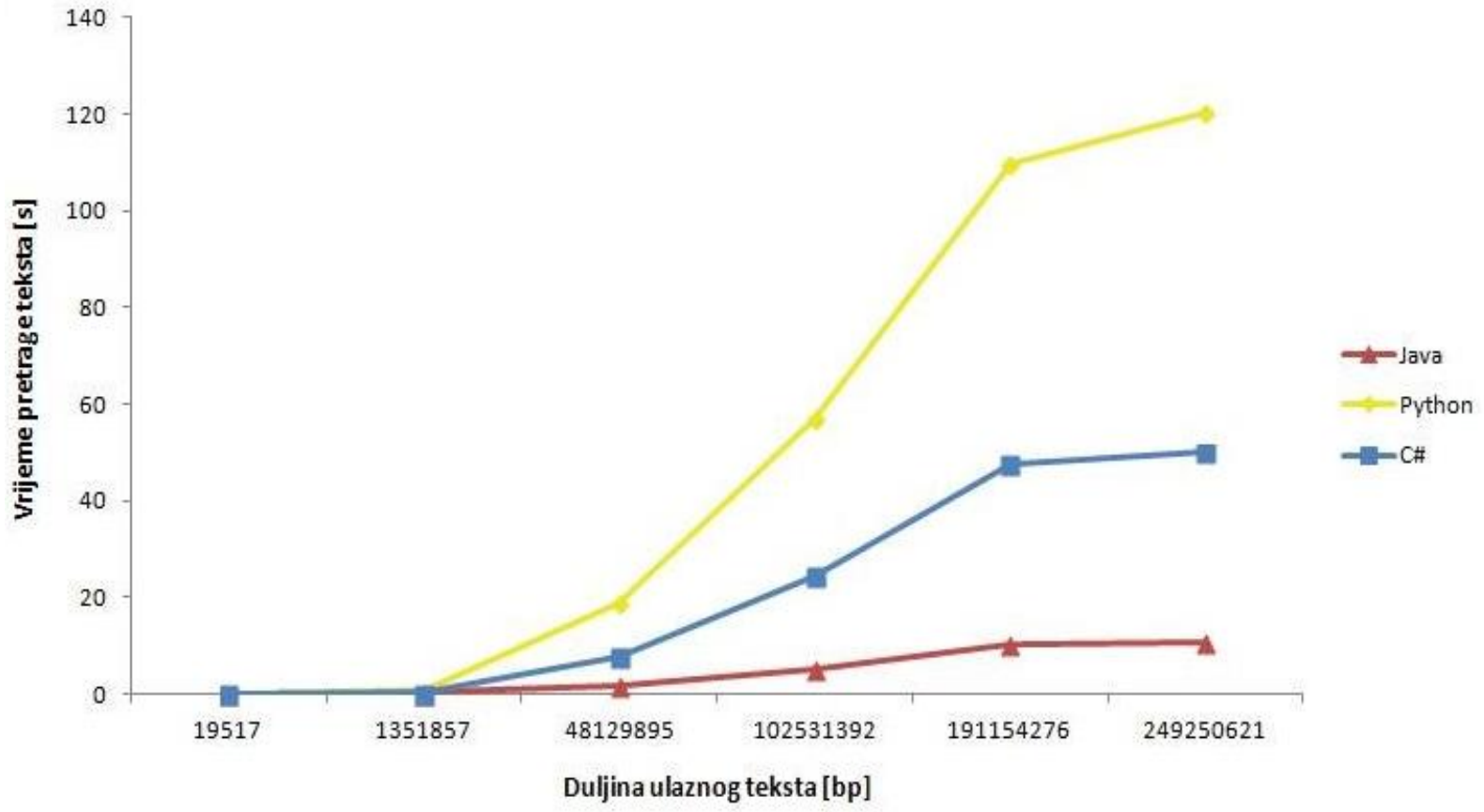
Implementacija i rezultati

a) vrijeme izgradnje DFA



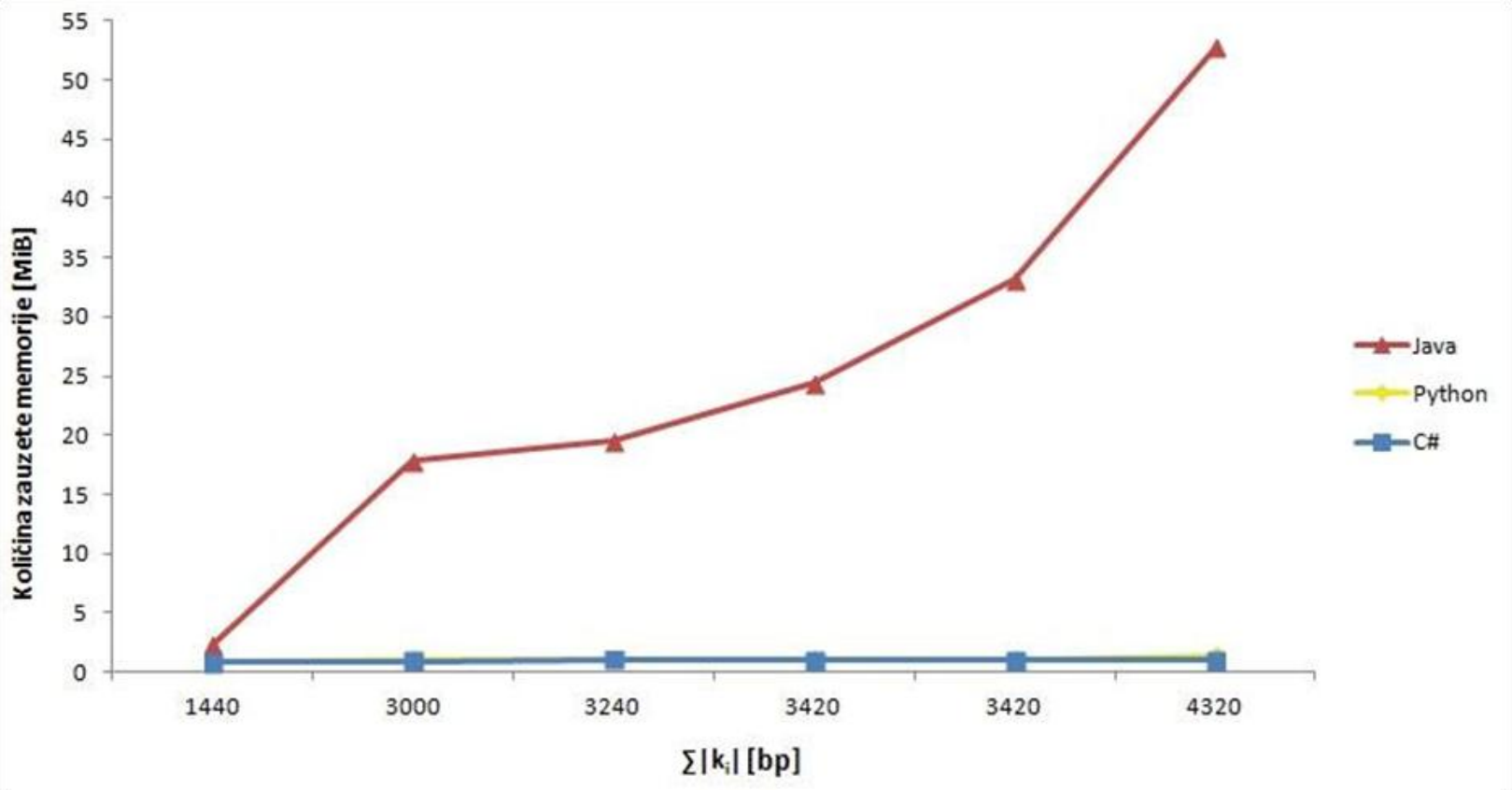
Implementacija i rezultati

b) vrijeme prolaska teksta kroz DFA

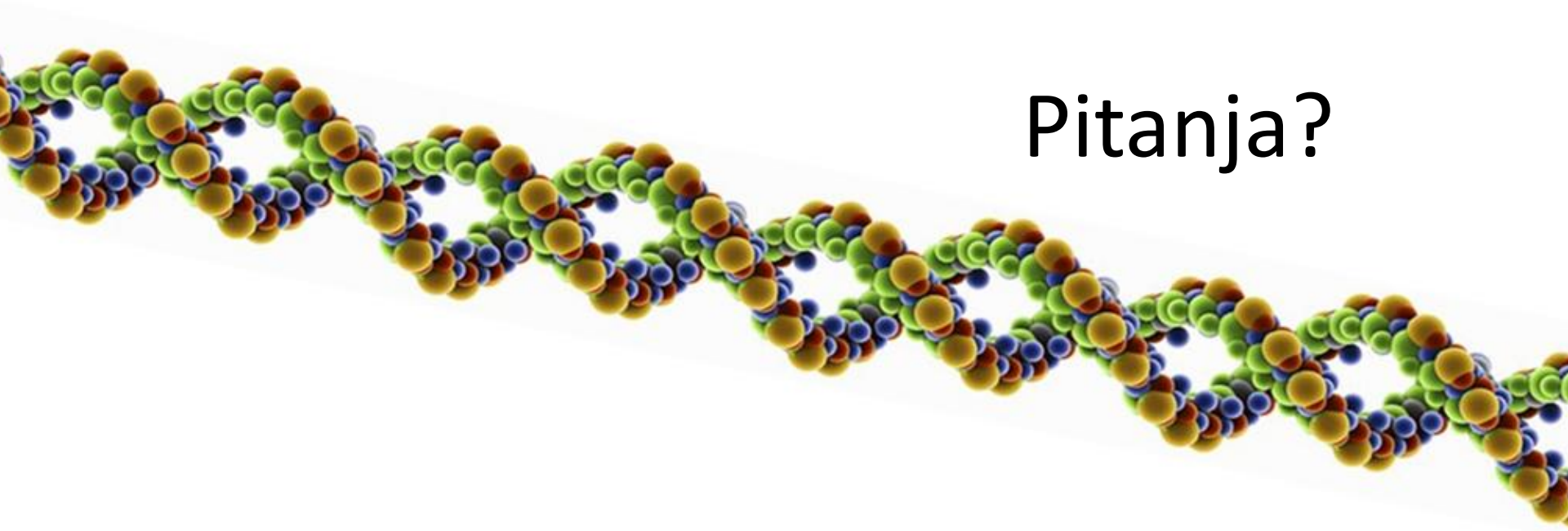


Implementacija i rezultati

c) utrošena memorija



Hvala na pozornosti ☺



Pitanja?