Litery, ogonki i emotikonki

Krótka opowieść o tym skąd komputer wie jaki znak ma wyświetlić

Festiwal Nauki, Techniki i Sztuki, Łódź, 12 maja 2022

Piotr Fulmański

piotr.fulmanski@wmii.uni.lodz.pl
piotr@fulmanski.pl
https://fulmanski.pl

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Jaki język rozumie komputer?

???

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Jaki język rozumie komputer?

Nasze komputery mają problemy ze zrozumieniem, jak homo sapiens mówi, czuje czy śni.

Yuval Noah Harari, "Sapiens. Od zwierząt do bogów"

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Jaki język rozumie komputer?

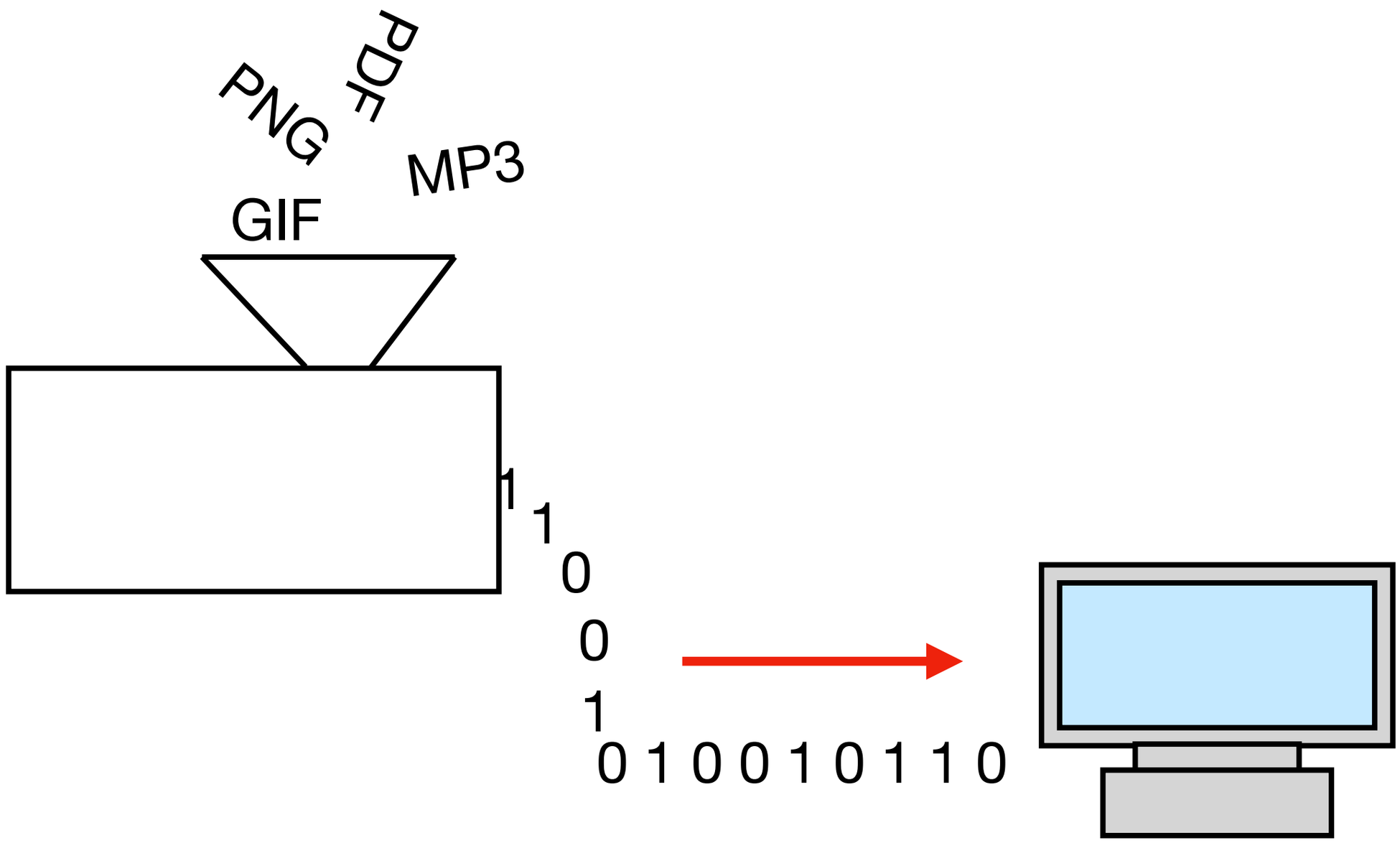
Nasze komputery mają problemy ze zrozumieniem, jak homo sapiens mówi, czuje czy śni. Uczymy więc homo sapiens mówić, czuć czy śnić w języku liczb, który jest zrozumiały dla komputerów.

Yuval Noah Harari, "Sapiens. Od zwierząt do bogów"

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Język jaki rozumie komputer

Tym językiem jest język liczb... Ciąg liczb... Ciąg cyfr *zero* i *jeden*.



WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Język jaki rozumie komputer

Wszystko co istnieje w komputerze musi dać się zapisać jako ciąg *zer* i *jedynek*.

```
01000010 01001101 10000110 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00110110 00000000 00000000 00000000 00101000 00000000
00000000 00000000 00000101 00000000 00000000 00000000 00000101 00000000
00000000 00000000 00000001 00000000 00011000 00000000 00000000 00000000
00000000 00000000 01010000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 11111111 11111111 11111111 11111111 11111111 00000000
11111111 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 11111111 11111111 11111111 11111111 11111111 11111111 11111111
11111111 11111111 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 11111111 11111111 11111111 11111111 11111111
11111111 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Język jaki rozumie komputer

Wszystko co istnieje w komputerze musi dać się zapisać jako ciąg *zer* i *jedynek*.

```
42 4d 86 00 00 00 00 00 00 00 36 00 00 00 28 00
00 00 05 00 00 00 05 00 00 00 01 00 18 00 00 00
00 00 50 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 ff ff ff ff ff 00
ff 00 00 00 00 00 00 00 00 ff ff ff ff ff ff ff
ff ff 00 00 00 00 00 00 00 00 00 ff ff ff ff ff
ff 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00
```

```
01000010 01001101 10000110 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00110110
00000000 00000000 00000000 00101000 00000000 00000000 00000000 00000101 00000000 00000000 00000000
00000101 00000000 00000000 00000000 00000001 00000000 00011000 00000000 00000000 00000000 00000000
00000000 01010000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 11111111 11111111 11111111
11111111 11111111 00000000 11111111 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 11111111 11111111 11111111 11111111 11111111 11111111 11111111 11111111 11111111 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 11111111 11111111 11111111
11111111 11111111 11111111 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
00000000 00000000
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Język jaki rozumie komputer

Nawet liczby zapisywane są w specjalny sposób

```
 12   = 00000000 00000000 00000000 00001100 unsigned int
+12   = 00000000 00000000 00000000 00001100 signed int
-12   = 11111111 11111111 11111111 11110100 signed int
+12.0 = 01000001 01000000 00000000 00000000 ieee 754 32b
-12.0 = 11000001 01000000 00000000 00000000 ieee 754 32b
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja liczb

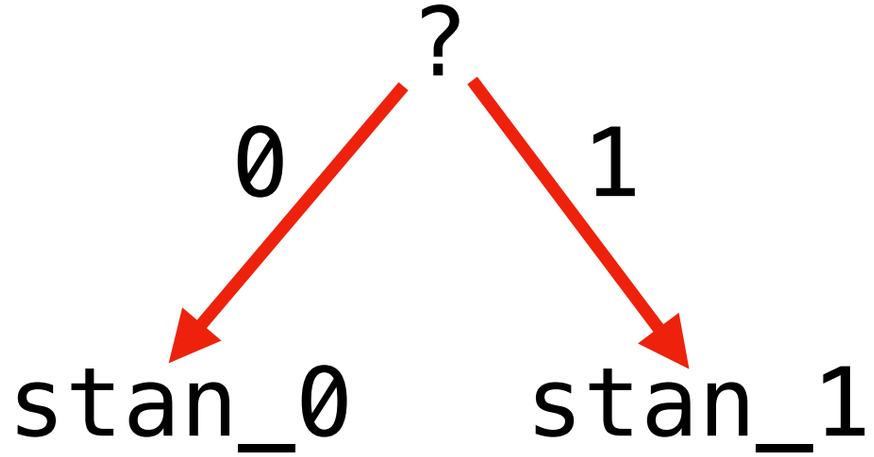
Ze względów praktycznych przywiązujemy wyjątkową wagę do reprezentacji liczby z zakresu 0-15 oraz 0-255.

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja liczb

Zakres 0-1

Bit – najmniejsza ilość informacji potrzebna do określenia, który z dwóch równie prawdopodobnych stanów przyjął układ.

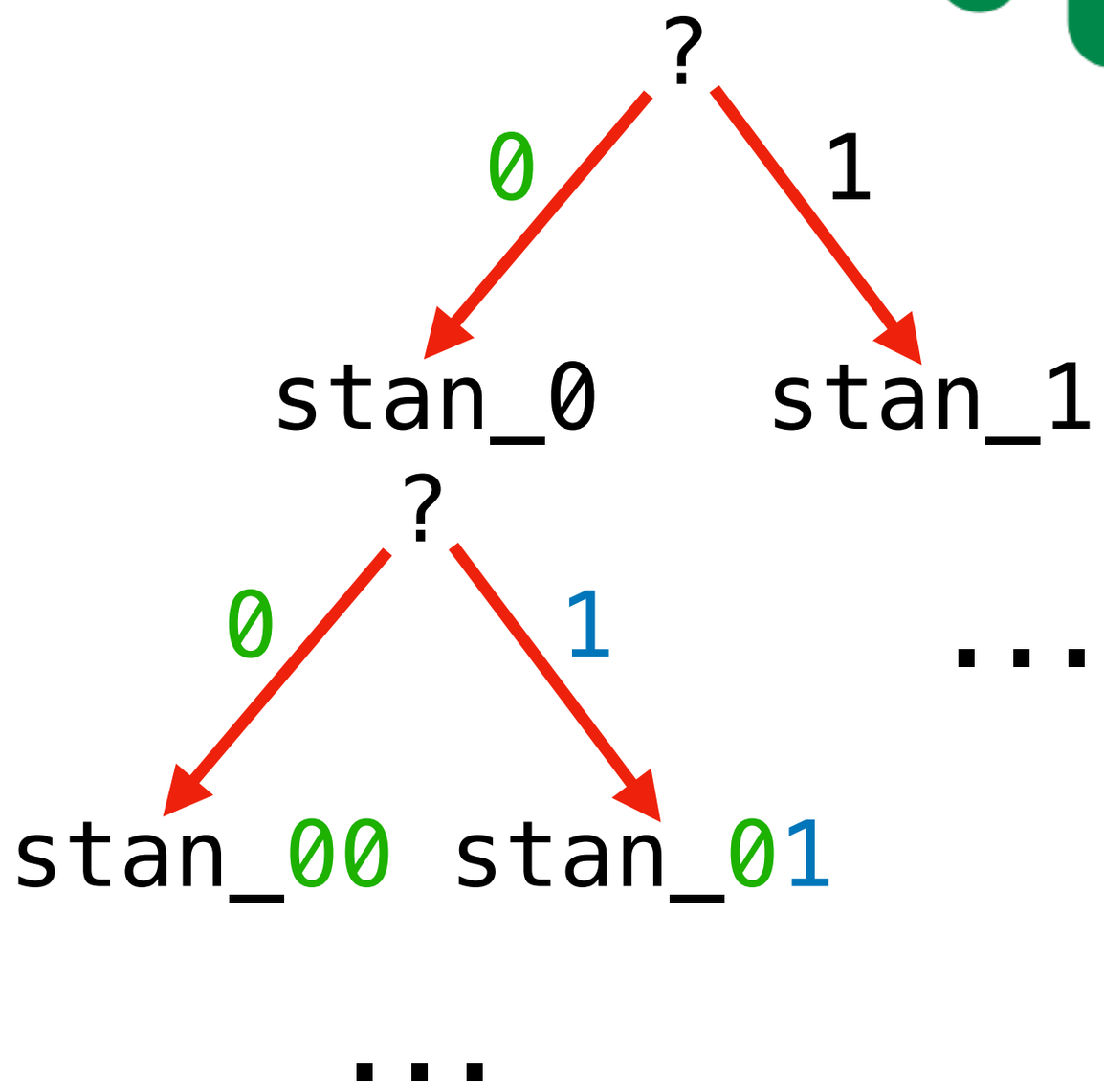


WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja liczb

Zakres 0-15

```
0 0000     8 1000
1 0001     9 1001
2 0010    10 1010
3 0011    11 1011
4 0100    12 1100
5 0101    13 1101
6 0110    14 1110
7 0111    15 1111
```



WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja liczb

Zakres 0-15

0	0000	0	8	1000	8
1	0001	1	9	1001	9
2	0010	2	10	1010	A
3	0011	3	11	1011	B
4	0100	4	12	1100	C
5	0101	5	13	1101	D
6	0110	6	14	1110	E
7	0111	7	15	1111	F

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja liczb

Zakres 0-255

Bajt – najmniejsza adresowalna jednostka informacji pamięci komputerowej.
Zwykle przyjmuje się, że jeden bajt to 8 bitów.

```
  0  00000000  00
  1  00000001  01
  2  00000010  02
 ...
253  11111101  FD
254  11111110  FE
255  11111111  FF
```

```
0 0000 0      8 1000 8
1 0001 1      9 1001 9
2 0010 2     10 1010 A
3 0011 3     11 1011 B
4 0100 4     12 1100 C
5 0101 5     13 1101 D
6 0110 6     14 1110 E
7 0111 7     15 1111 F
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja znaków
Zdroworozsądkowa

Czyli jaka?

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja znaków
Zdroworozsądkowa

```
A 1
B 110
C 10
D 101
E 011
F 100
G 11
...
Z 00
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja znaków
Zdroworozsądkowa

```
A 1
B 110
C 10
D 101
E 011
F 100
G 11
...
Z 00
```

```
101100 = D F
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja znaków
Zdroworozsądkowa

```
A 1
B 110
C 10
D 101
E 011
F 100
G 11
...
Z 00
```

```
101100 = D F

101100 = 10 11 00 = C G Z
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja znaków
Zdroworozsądkowa

```
A 1
B 110
C 10
D 101
E 011
F 100
G 11
...
Z 00
```

```
101100 = D F

101100 = 10 11 00 = C G Z

101100 = 1 011 00 = A E Z
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja znaków

Zdroworozsądkowa

- Wszystkie kody mają taką samą długość.
- Powinny być skonstruowane według jakiejś zasady ułatwiającej powiązanie ciągu 0-1 ze znakiem.
- W alfabecie łacińskim mamy 26 liter.
- Potrzebujemy zatem 52 rożnych kodów, na co potrzeba minimum 6 bitów.

Reprezentacja znaków
Zdroworozsądkowa

Tadam!!!

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Reprezentacja znaków

Zdroworozsądkowa

ASCII

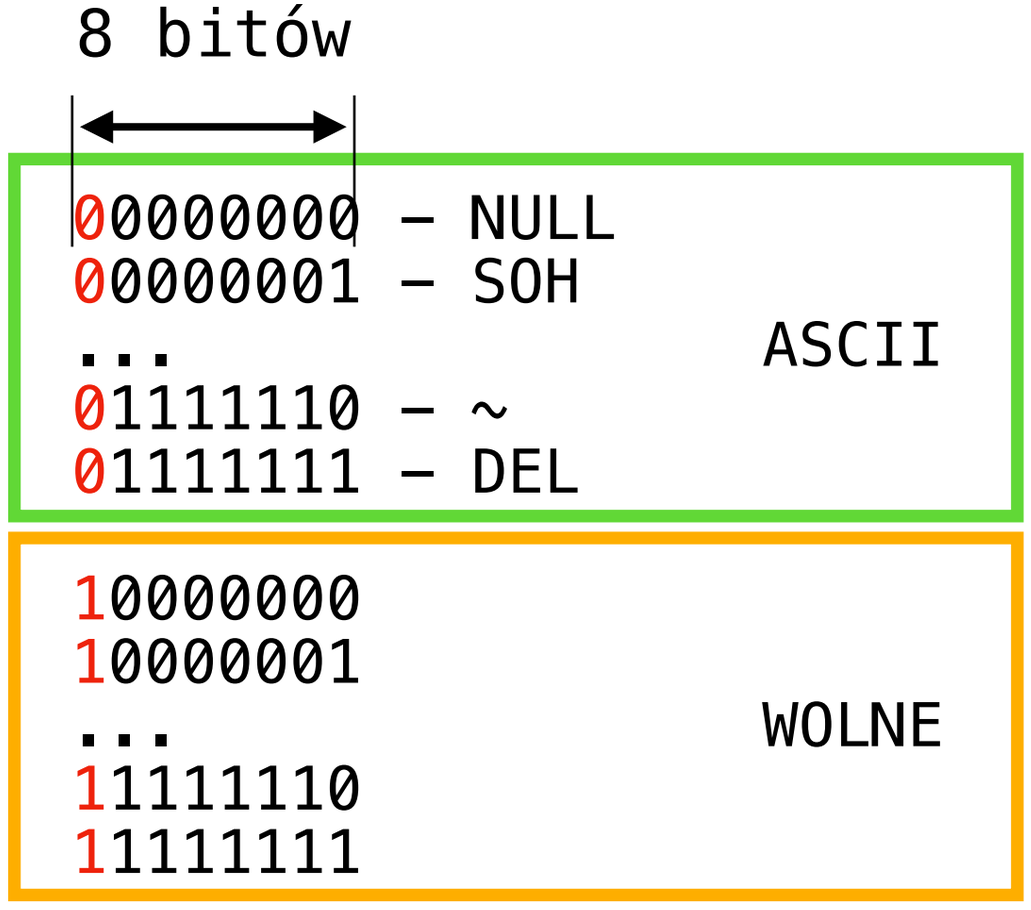
Tabela kodów ASCII

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

ASCII

Przyglądając się tabeli kodów ASCII widać, że kody kończą się na 127 a więc potrzebujemy 7 bitów. Dla wygody przyjmujemy kodowanie które do 7 bitów dodaje bit zerowy (a dokładnie to poprzedza je).

Dzięki temu, w ASCII jeden znak to zawsze jeden bajt, choć z 255 numerów wykorzystujemy jedynie 128.



WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

ASCII

Jakiego problemu ASCII nie rozwiązuje?

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

ASCII wielojęzyczne

```
0000000 – NULL
0000001 – SOH
                    ASCII
...
0111110 – ~
0111111 – DEL
```

```
10000000    Polski
10000001
...  Norweski      WOLNE
11111110
11111111    Ukraiński
```

Niemiecki

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

ASCII wielojęzyczne

Strony kodowe

```
00000000 – NULL
00000001 – SOH
...                    ASCII
01111110 – ~
01111111 – DEL
```

```
10000000
10000001
...
11111110
11111111
```
Polski

```
10000000
10000001
...
11111110
11111111
```
Niemiecki

```
10000000
10000001
...
11111110
11111111
```
Norweski

```
10000000
10000001
...
11111110
11111111
```
Ukraiński

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

ASCII wielojęzyczne

Strony kodowe

```
00000000 – NULL
00000001 – SOH
...                    ASCII
01111110 – ~
01111111 – DEL
```

```
10000000
10000001
...
11111110
11111111
```

Polski
ISO 8859-2

```
10000000
10000001
...
11111110
11111111
```

Polski
CP 1250

```
10000000
10000001
...
11111110
11111111
```

Polski
Mazovia

```
10000000
10000001
...
11111110
11111111
```

Polski
Calamus

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Kodowanie polskich znaków

ASCII wielojęzyczne

Strony kodowe

Aaaaaaaaaaaaa!

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Unicode na ratunek

Unicode na ratunek

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Unicode na ratunek

Unicode to system przypisujący określonemu symbolowi, określony numer

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Unicode a UTF

Zupełnie inną kwestią jest, w jaki sposób te numery Unicode będą zakodowane w ciągach składających się z zer i jedynek.
Jedną z możliwości jest kodowanie UTF (*Unicode Transformation Format*).

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
UTF-8

Mapowanie znaków Unicode na ciągi bajtów w kodowaniu UTF-8:

KODY	BAJTY
0x00 do 0x7F	– 0xxxxxxx
0x80 do 0x7FF	– 110xxxxx 10xxxxxx
0x800 do 0xFFFF	– 1110xxxx 10xxxxxx 10xxxxxx
0x10000 do 0x1FFFFF	– 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
0x200000 do 0x3FFFFFF	– 111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
0x4000000 do 0x7FFFFFFF	– 1111110x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

UTF-8

```
$ echo -n -e '\xc4\x84' > unknown_utf8.txt
$ od -t x1 unknown_utf8.txt
0000000 c4 84
0000002
$ cat unknown_utf8.txt
Ą
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Zanim jednak zabierzmy się za kodowanie w UTF-8, to musimy dokładnie wiedzieć co zakodować i w jakiej kolejności. A to wcale nie jest takie oczywiste...

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

ą czyli *a z ogonkiem*

ą to według Unicode znak o kodzie (szesnastkowym) 0105
nazywany: LATIN SMALL LETTER A WITH OGONEK kod

Liczba 105 jest kodowana UTF-8 za pomocą **dwóch bajtów**: C4 85.

```
$ echo -n -e '\xc4\x85' > unknown_utf8.txt
$ cat unknown_utf8.txt
ą
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

ą czyli *a* z *ogonkiem*

a to według Unicode znak o kodzie (szesnastkowym) 61
nazywany: LATIN SMALL LETTER A kod

"ogonek" to według Unicode znak o kodzie (szesnastkowym) 328
nazywany: COMBINING OGONEK kod

61 w UTF-8 to 61, 328 w UTF-8 to CC A8
W tym przypadku litera ą jest kombinacją (złożeniem) tych dwóch znaków, kodowaną na **trzech bajtach**:

```
$ echo -n -e '\x61\xcc\xa8' > unknown_utf8.txt
$ cat unknown_utf8.txt
ą
$ echo -n -e '\xcc\xa8\x61' > unknown_utf8.txt
$ cat unknown_utf8.txt
̨a
```

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Litera?

Co jest dla nas najmniejszą, niepodzielną częścią zdania
(w sensie jego "graficznego" zapisu)?

Litera?

Znak?

A może coś innego?

'a' to ...

'ą' to ...

'ż' to...

'rz' to ...

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Grafem i *glif*

Grafem – (lub symbol) najmniejsze jednostka pisma

Glif – kształt przedstawiający w określonym kroju pisma konkretny grafem lub symbol

W przypadku 'r' i 'z' tworzących 'rz' nie jest to takie widoczne.
Jaśniejsze staje się w przypadku 'a' i 'ogonka' tworzących 'ą' – ogonek, jako samodzielny znak, "nie istnieje".

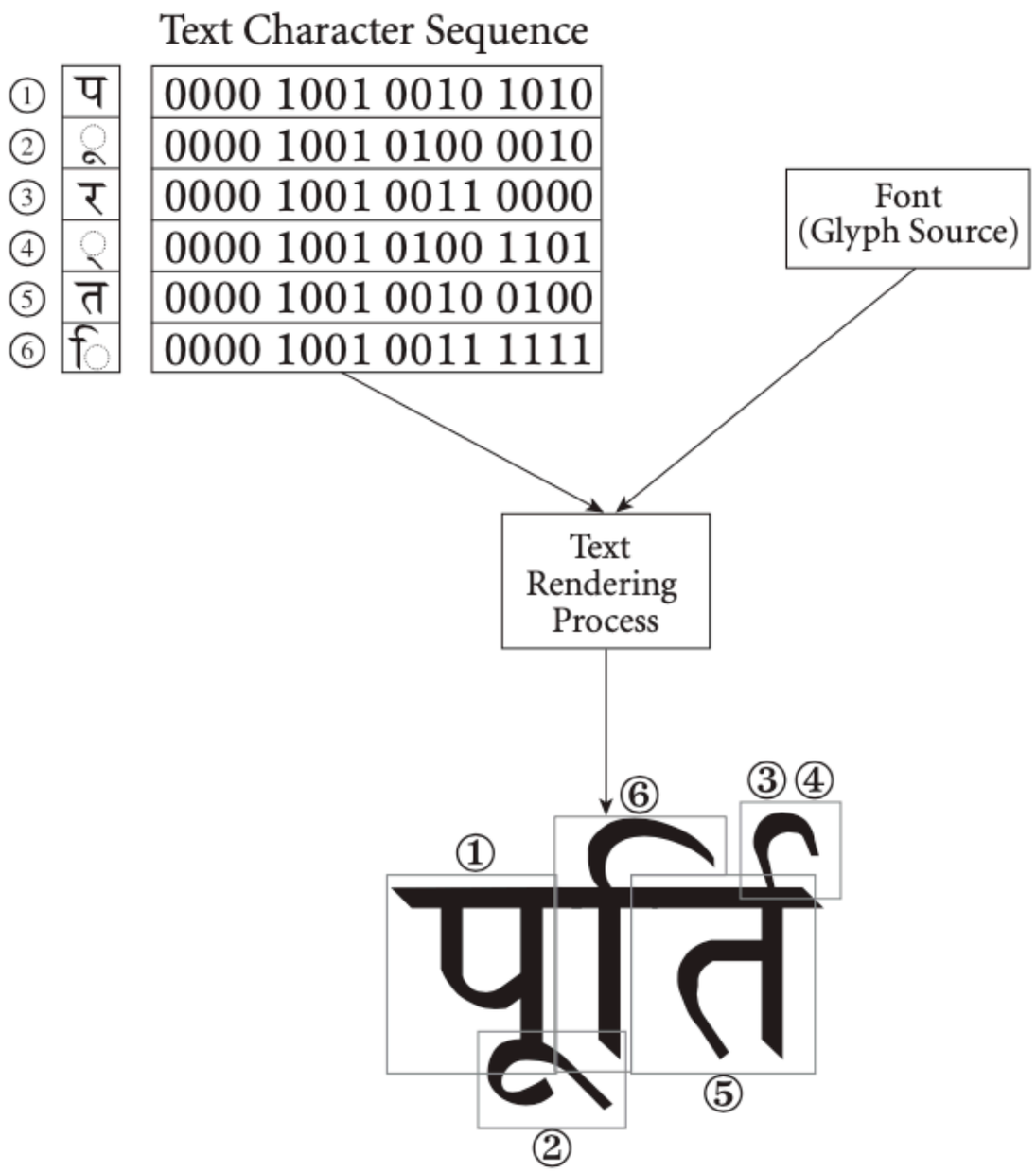
Ogonek jest pewnego rodzaju "modyfikatorem" (tzw. *combining mark*).

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Combining marks w działaniu

Figure 2-3. Unicode Character Code to Rendered Glyphs



WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Obrazek pochodzi z: The Unicode® Standard Version 12.0, March 2019

Unicode
Emotikonki

Niektóre, jak "normalne" znaki kodowane są w "naturalny" sposób.

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
Emotikonki

Niektóre, jak "normalne" znaki kodowane są w "naturalny" sposób.

SMILING FACE WITH SMILING EYES

kod: U+1F60A

UTF-8: F0 9F 98 8A

([1], [2])

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
Emotikonki

Inne zaś wymagają użycia modyfikatorów.

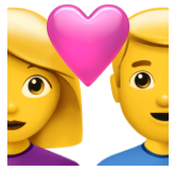
Unicode
Emotikonki

Couple with Heart: Woman, Man ([1])

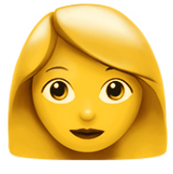
kod: brak jednego kodu

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
Emotikonki



Couple with Heart: Woman, Man

 WOMAN, U+1F469

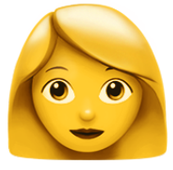
UTF-8: F0 9F 91 A9

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Emotikonki

Couple with Heart: Woman, Man



WOMAN, U+1F469

HEAVY BLACK HEART, U+2764

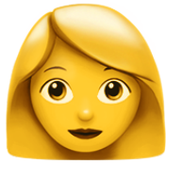
UTF-8: F0 9F 91 A9 E2 9D A4

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

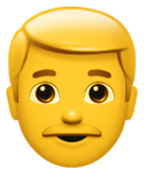
Emotikonki

Couple with Heart: Woman, Man



WOMAN, U+1F469

HEAVY BLACK HEART, U+2764



MAN, U+1F468

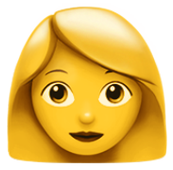
UTF-8: F0 9F 91 A9 E2 9D A4 F0 9F 91 A8

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Emotikonki

Couple with Heart: Woman, Man

 WOMAN, U+1F469

ZERO WIDTH JOINER, U+200D

 HEAVY BLACK HEART, U+2764

ZERO WIDTH JOINER, U+200D

 MAN, U+1F468

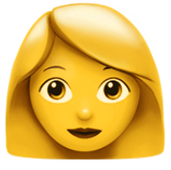
UTF-8: F0 9F 91 A9 E2 80 8D E2 9D A4 E2 80 8D F0 9F 91 A8

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Emotikonki

Couple with Heart: Woman, Man



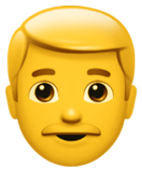
WOMAN, U+1F469

ZERO WIDTH JOINER, U+200D

HEAVY BLACK HEART, U+2764

VARIATION SELECTOR-16, U+FE0F

ZERO WIDTH JOINER, U+200D



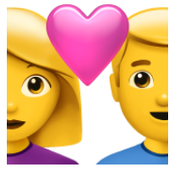
MAN, U+1F468

UTF-8: F0 9F 91 A9 E2 80 8D E2 9D A4 EF B8 8F E2 80 8D F0 9F 91 A8

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Emotikonki



Couple with Heart: Woman, Man

Unicode: U+1F469 U+200D U+2764 U+FE0F U+200D U+1F468

UTF-8: F0 9F 91 A9 E2 80 8D E2 9D A4 EF B8 8F E2 80 8D F0 9F 91 A8

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Emotikonki

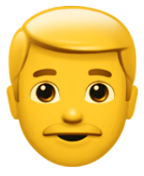
Firefighter: Dark Skin Tone ([1])

kod: brak jednego kodu

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
Emotikonki

Firefighter: Dark Skin Tone

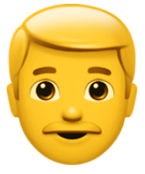
 MAN, U+1F468

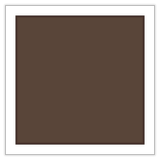
UTF: F0 9F 91 A8

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
Emotikonki

Firefighter: Dark Skin Tone

 MAN, U+1F468

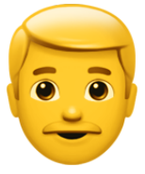
 EMOJI MODIFIER FITZPATRICK TYPE-6, U+1F3FF

UTF-8: F0 9F 91 A8 F0 9F 8F BF

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
Emotikonki

Firefighter: Dark Skin Tone

 MAN, U+1F468

 EMOJI MODIFIER FITZPATRICK TYPE-6, U+1F3FF

ZERO WIDTH JOINER, U+200D

UTF-8: F0 9F 91 A8 F0 9F 8F BF E2 80 8D

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode
Emotikonki

Firefighter: Dark Skin Tone

 MAN, U+1F468

 EMOJI MODIFIER FITZPATRICK TYPE-6, U+1F3FF

ZERO WIDTH JOINER, U+200D

 FIRE ENGINE, U+1F692

UTF-8: F0 9F 91 A8 F0 9F 8F BF E2 80 8D F0 9F 9A 92

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Unicode

Emotikonki

Firefighter: Dark Skin Tone

Unicode: U+1F468 U+1F3FF U+200D U+1F692

UTF-8: F0 9F 91 A8 F0 9F 8F BF E2 80 8D F0 9F 9A 92

WYDZIAŁ
MATEMATYKI
i INFORMATYKI
Uniwersytet Łódzki

Inne materiały

Template normal

Template normal

- https://emojipedia.org
- Full Emoji List, v14.0, https://www.unicode.org/emoji/charts/full-emoji-list.html
- Text encodings, https://fulmanski.pl/tutorials/computer-science/computer-science-for-beginners/text-encodings/