12. Kódování znaků

1. ASCII

- American Standard Code for Information Interchange
- definuje znaky zejména používané v informatice
- obsahuje sedmibitové znaky, tedy celkom 128 znaků
- 0 31 jsou netisknutelné znaky, jsou k řízení datového přenosu, formátování tisku
- spravuje organizace ISO

2. Unicode

- konzistentní znaková sada
- reprezentuje se v ní více než 14000 znaků používaných po celém světě
- pokud chceme zakódovat určitý znak do unicodu, použijeme unicode tabulku
 - -ale jednoduše latinka začíná 97 a, a končí 122 z
 - potom převedeme decimální číslo na hex, takže 97 by bylo 61
 - unicode se zapisuje U+hex, pro nás U+0061
- principy kódování:
 - jednotnost: konstantní šířka dovoluje rychlé hledání, třídění, ...
 - univerzálnost: zahrnuje všechny znaky, které by mohly být použité při výměně textu
 - jednoznačnost
 - maximální využití: není nutná escape sekvence, znak není závislý na jeho kontextu, snadaná zpracovávatelnost strojem

3. UTF-8

- unicode transformation format
- jeden ze způsobů kódování, znaky na číselné řetezce
- má proměnnou délku od 1 do 4 bajtů
- vychází ze standardu Unicode
- ostatní, například UTF-16 či UTF-32 mají fixní délku, 16 respektive 32 bitů
- zpětná kompatabilita s ASCII

4. Huffmanovo kódování

- bezeztrátová komprese dat
- základní princip spočívá v tom, že se znaky, které se v souboru vyskytují nejčastěji, jsou konvertovány do řetězců s nejkratší délkou, nejfrekventovanější znak může být konvertován do jediného bitu
- komprese probíhá ve dvou krocích, řazení dle četnosti jednotlivého znaku, potom vytvoření binárního stromu
- postup při kódování ABRAKADABRA:

5. Shannon - Fano kódování

- jedná se o bezeztrátovou kompresi dat
- od Huffmanova kódování se liší konstrukcí binárního stromu

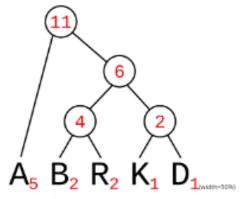
$A_5 B_2 R_2 K_1 D_1$

Figure 1: Seřazení dle četnosti jednotlivých znaků - Huffman

- množina znaků je dělena na dvě, tak aby součet znaků v každé byl připližně stejný
- poté je první přiřazena 0 a druhé 1

Vytvoření stromu

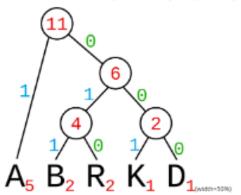
K vytvoření stromu je třeba spojit dva objekty s rejmenším počtem výskytů a jejich Pokud jsou síce než dva objekty se stejným číslem, můžeme si šloovoině vytrat.



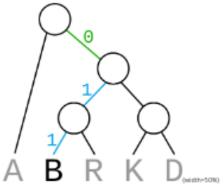
Každá objekt může mít pouze dva potorsky. Tím je strom kompřetní

Zakódování vstupu

Každou větev stromu si označíme, například levou větev jako 1 a prasou větev jako



Zakódování pak probíhá jen tak, že zmapují cestu ke znaku.



Draky budou mit tedy tyto ködy

A 1 H 811 K 815 K 815

Zakódování řetiloce už je jen o popsání správných jedniček a nul podle vstupu

ARREAADERA - 10150185001180038118381

Figure 2: Vytvoříme strom a provedeme následující kroky - Huffman $\stackrel{\cdot}{3}$

Symbol	Α	В	С	D	Е
Probabilities	0.385	0.179	0.154	0.154	0.128
First division	0		1		
Second division	0	1	0	1	
Third division				0	1
Codewords	00	01	10	110	111

Figure 3: Tabulka zobrazující strom a jednotlivé kódové řetězce - Shannon