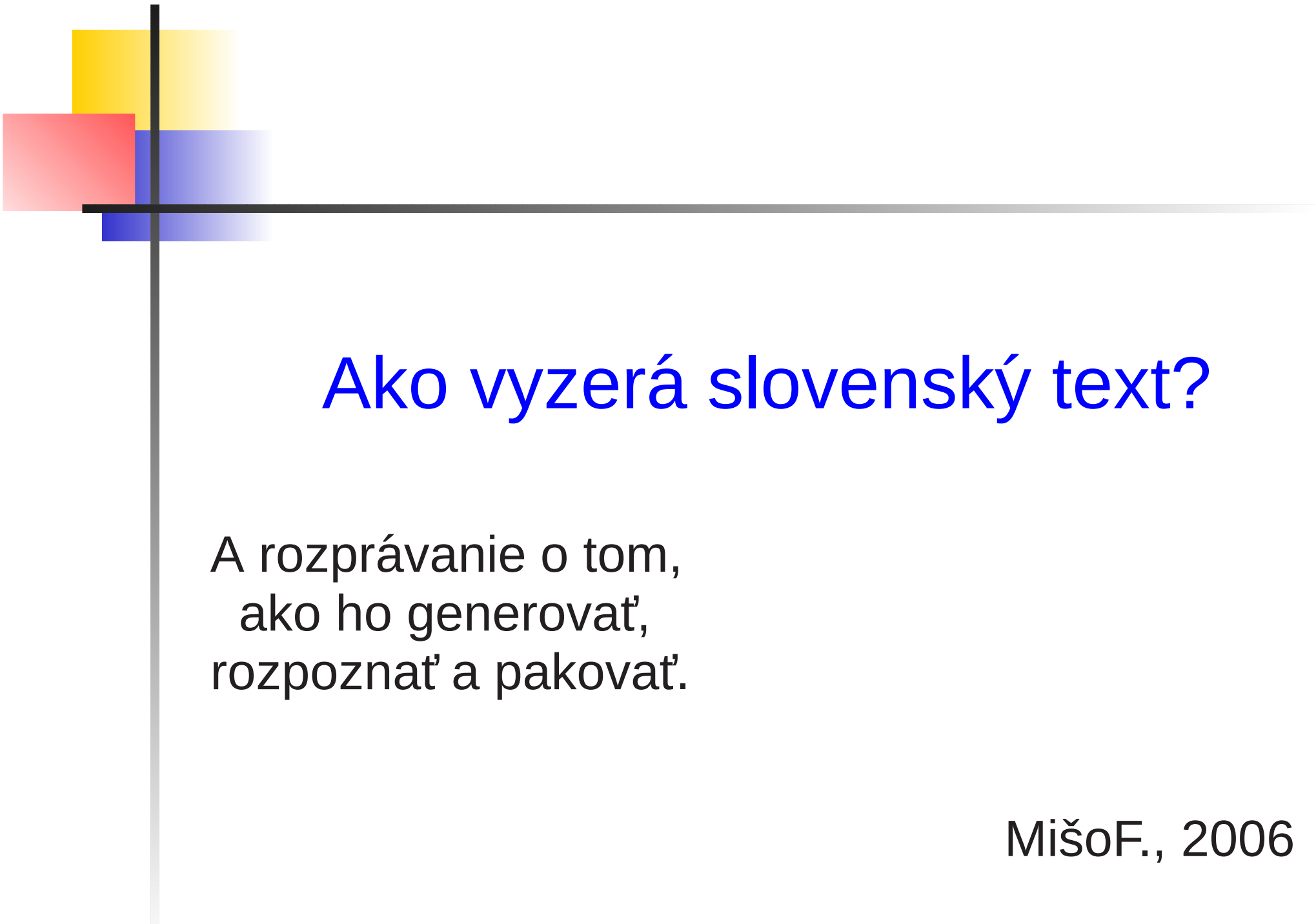
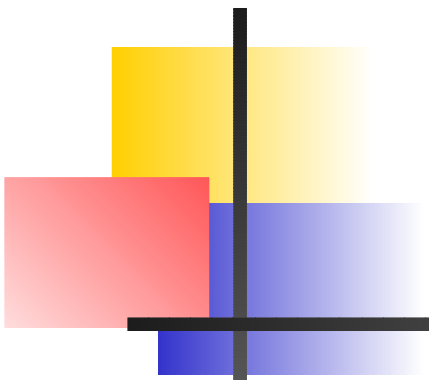


Ako vyzerá slovenský text?

A rozprávanie o tom,
ako ho generovať,
rozpoznať a pakovať.

MišoF., 2006



Generovanie

Prvý pokus

Len tak náhodné znaky:

```
W TVJQDIDSIDWRIRGXVIRUNNMWWACU
AYIIIGYLPRSMUDCSJYDTFUMIGNDRDF
ADSYA EYAIDHUN XEJUXBPFDMBQEHI
ZXLFVLVZZKWBRPOGAHPKTFOZY  OTW
LRTMLNMFAKPWBW EBABF U NSYBGAU
```

Generovanie

Prvý pokus

Toto slovenčina nie je.

W TVJQDIDSIDWRIRGXVIRUNNMWWACU
AYIIIGYLPRSMUDCSJYDTFUMIGNDRDF
ADSYA EYAIDHUN XEJUXBPFDMBQEHI
ZXLFVLVZZKWBRPOGAHPKTFOZY OTW
LRTMLNMFAKPWBW EBABF U NSYBGAU

(A kde sú medzery?)

Generovanie

Druhý pokus

Slovenčina má viac A ako X

a	9.31%	h	1.93%	o	7.54%	v	3.34%
b	1.32%	i	5.98%	p	2.64%	w	0.06%
c	2.84%	j	1.40%	q	0.01%	x	0.10%
d	2.98%	k	3.39%	r	4.40%	y	2.09%
e	7.63%	l	3.74%	s	4.81%	z	2.25%
f	0.43%	m	2.92%	t	4.91%	_	15.56%
g	0.45%	n	5.07%	u	2.91%		

Generovanie

Druhý pokus

Použijeme tieto pravdepodobnosti

```
A OMHATRNOPNVNBANOVOAEPHITNFLI
INCH  WA IPOEZNIL IE UL NDS  A
SRAI LDDNP OPC A HMSNYLIW   IZ
O  OUI LMKMO UAZAON XKHRHTAA O
O AIENSCZCNHTLNHTNGEMAODCIZDZI
```

Generovanie

Druhý pokus

Lepšie? O moc nie. . .

```
A OMHATRNOPNVNBANOVOAEPHITNFLI
INCH  WA IPOEZNIL IE UL NDS  A
SRAI LDDNP OPC A HMSNYLIW???IZ
O  OUI LMKMO UAZAON XKHRHTAA O
O AIENSCZCNHTLNHTNGEMAODCIZDZI
```

. . . a čítať sa to teda nedá.

Generovanie

Prečo sa to nedá čítať?

Ignorujeme závislosti medzi písmenami.
A čo tie hovoria?

- CH oveľa častejšie ako FH
- anglické QU
- striedanie samohlások/spoluhlások
- nechceme dve medzery za sebou

Generovanie

Tretí pokus

Pravdepodobnosť závisí od minulého písmena

```
AM MIERE HLIE UDIEDEJERA JUM Z
 PEKTORSKOVO ZIAJMICOPS BRATA
U O PRA PA ODODOL SM OVNIE OPR
NASEPE TU NCETVI CHUDNEGTISICK
ORAZI M ZA VENOLTALERAM NK DIL
```

Generovanie

Tretí pokus

Lepšie, ale stále nie ono.

AM MIERE HLIE UDIEDEJERA JUM Z
PEKTORSKOVO ZIAJMICOPS BRATA
U O PRA PA ODODOL SM OVNIE OPR
NASEPE TU NCETVI CHUDNEGTISICK
ORAZI M ZA VENOLTALERAM NK DIL

Generovanie

Štvrtý pokus

Tak nech písmeno závisí od predchádzajúcej trojice.

Čo by ste doplnili za . . .

- GRA
- SLO
- NEH
- PRE

Generovanie

Štvrtý pokus

Tak nech písmeno závisí od predchádzajúcej trojice.

Čo by ste doplnili za . . .

- GRA: GRAF, PROGRAM, GRANULE
- SLO: SLOVO, SLOVAK, HESLO_
- NEH: . . . NEHO, NEHA, SNEH_
- PRE: PRE_, PRED. . . , PRES. . .

Generovanie

Štvrtý pokus

Toto z toho vylezie:

```
POLOIDU DUO MAJSTERAZ O URCIAC
LUDIL PRED NOVY SUBMI DOSTRETU
 V TENT SME ROHO VIAC SU A KOJ
ACUL SA NASOCID ROBCHOL BYTOVA
 TROMY VLADE THAILO BLOVANIE A
```

použili sme pp výskytu *tetragramov*.

Zbieranie dát

Menej významná otázka:

Kde zohnať dáta?

Nepotrebujeme presné(?!) %.
Stačí nám *dosť veľká* vzorka textu.

Vyberáme ďalšie písmeno za XYZ:

- nájdeme náhodný výskyt XYZ v známom texte
- vyberieme písmeno za ním

Zbieranie dát

Menej významná otázka:

Kde zohnať dáta?

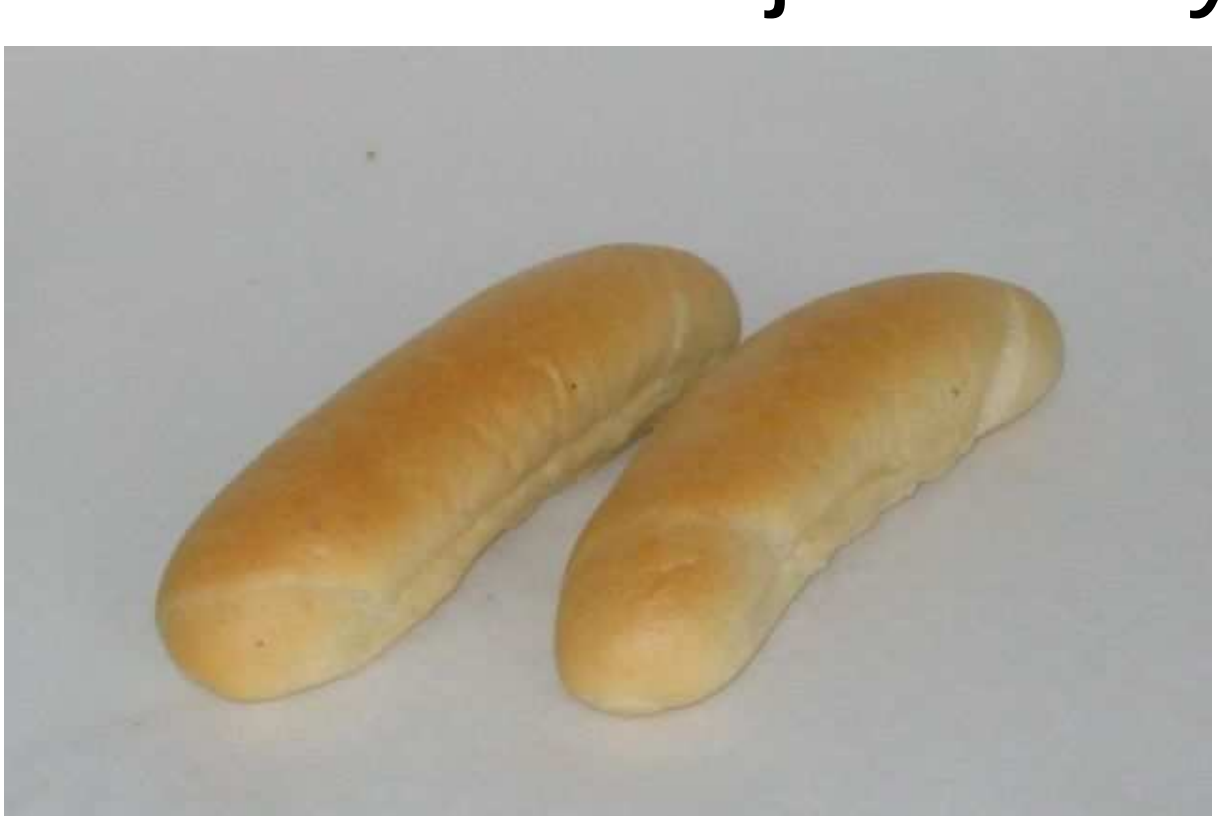
Vzorka? Nič ľahšie!

```
wget -r -np -k http://sme.sk/
for f in ‘find . -name ’*.html’‘ ; do
  lynx -dump --force-html ”$f” >> text
done
```

A čo s tým

Významnejšia otázka:

Budú teraz lacnejšie rožky?



A čo s tým

Významnejšia otázka:

Alebo aspoň treska?

A čo s tým

Významnejšia otázka menej poeticky:

Je nám to na niečo dobré?

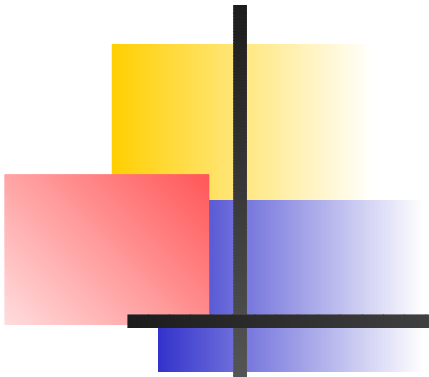
A čo s tým

Významnejšia otázka menej poeticky:

Je nám to na niečo dobré?

Na veľa vecí:

- rozpoznanie jazyka
- automatické dešifrovanie
- kompresia textu
- *bohužiaľ nie lacnejšia treska*



Pravdepodobnosť

- koľko treba hodov, kým padne znak?
- koľko treba hodov, kým padne šestka?

Pravdepodobnosť

- koľko treba hodov, kým padne znak?
- koľko treba hodov, kým padne šestka?
- formálne:

$$S = \sum_{i=1}^{\infty} (1-p)^{i-1} pi = ?$$

Pravdepodobnosť

- koľko treba hodov, kým padne znak?
- koľko treba hodov, kým padne šestka?
- formálne:

$$
\begin{aligned}
S &= \sum_{i=1}^{\infty}(1-p)^{i-1}pi = \\
&= \left(\sum_{i=1}^{\infty}(1-p)^{i-1}p\right) + \left(\sum_{i=1}^{\infty}(1-p)^{i-1}p(i-1)\right) = \\
&= 1 + (1-p)S \\
S &= \frac{1}{p}
\end{aligned}
$$

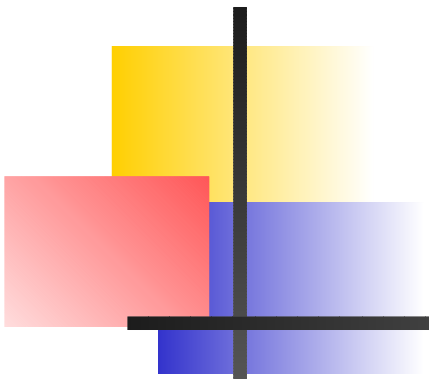
Pravdepodobnosť

Prekvapenie z toho, že jav J s pp. p nastal:

$$Q(J) = \lg \frac{1}{p}$$

Príklady a vysvetlenie:

- Čím $<$ pravdep. jav, tým $>$ prekvapenie.
- Pre nezávislé javy sa prekvapenie sčítava.
- Hádžeme k mincami naraz.
  Prekvapenie z toho, že padli samé znaky, je k.



Pravdepodobnosť

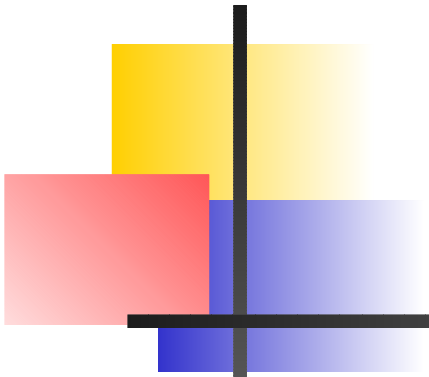
Akú jednotku má prekvapenie?

Pravdepodobnosť

Akú jednotku má prekvapenie?

BIT !

- Nastanie javu nám dáva informáciu.
- Čím menej pp jav, tým viac informácie.
- Príklad: náhodný deň roku 2006.
  Je to štátny sviatok.
  Dostali sme veľa informácie.



V akom som jazyku?

Je `THE DUCK SAID QUACK` po slovensky?
Ak nie, v akom jazyku to je?
Ako na to vie počítač prísť?

- Slovníky? Čo ak nemám nikde `QUACK`?
- Prekvapenie.
  V slovenskom texte je veľmi prekvapivé stretnúť tetragram `QUAC`.

V akom som jazyku?

Prekvapenie z tetragramu (v danom jazyku):

$$Q(xyzw) = \lg 1/pp(\text{náh. tetragram je } xyzw)$$

$$\approx \lg \frac{\text{dĺžka známeho textu}}{\text{počet } xyzw \text{ v známom texte}}$$

Prekvapenie z textu (v danom jazyku):
súčet prekvapenia z jeho tetragramov

V akom som jazyku?

angličtina:

```
pray take a seat said holmes this is my
friend and colleague dr watson who is oc
casionally good enough to help me in my
```

2264.91

čeština:

```
to bylo tenkrat kdyz pejsek a kocicka je
ste spolu hospodarili meli u lesa svuj m
aly domecek a tam spolu bydleli a chteli
```

1862.23

slovenčina (Botto):

```
kukala kukucka z zeleneho bucka sedem ra
z skukala sedem rockov dala janicko jank
o nas biele uz licka mas jasne uz ocka m
```

1646.13

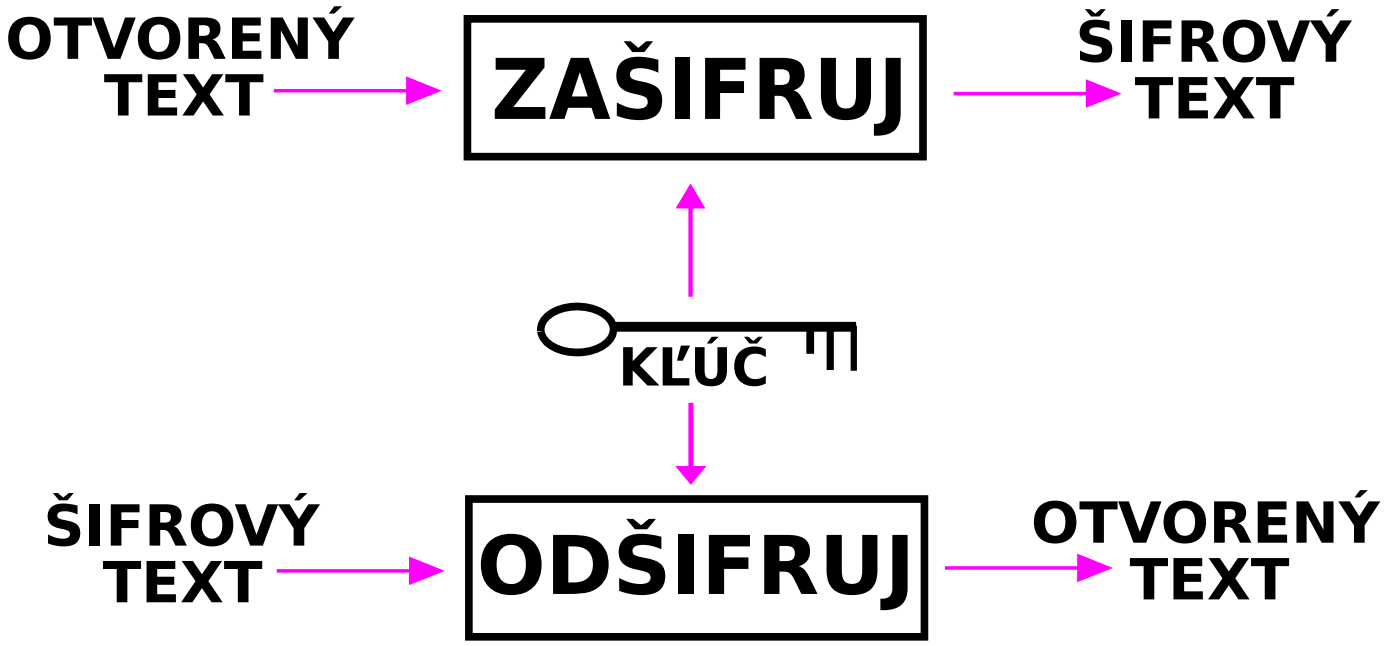
slovenčina (SME):

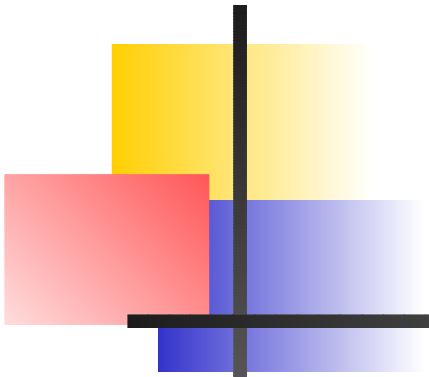
```
jeden kubicky meter zemneho plynu dokaze
 pri svojom spaleni vyrobit iste mnozstv
o energie za idealnych podmienok nam ten
```

1613.05

Šifry

Princíp symetrickej šifry:





Lámanie šifry

- máme *šifrový text*
- skúšame možné kľúče
- dorátame zodpovedajúce otvorené texty
- vyberieme len rozumne znejúce
- hodíme na výstup
- nejde to lepšie?

Lámanie šifry

Nejde to lepšie?

Môže, ale nie zadarmo. Čo treba?

lokálna (malá) zmena kľúča

$\longrightarrow$

lokálna (malá) zmena *otvoreného textu*

Substitučná šifra

Princíp: prehádžeme písmená
napr. namiesto každého a dáme x

Kľúč: permutácia množiny $\{a, \ldots, z\}$

Príklad:

Kľúč:

ABCDEFGHIJKLMNOPQRSTUVWXYZ

QWERTYUIOPASDFGHJKLZXCVBNM

Otvorený text:

JEDEN KUBICKY METER ZEMNEHO PLYNU DOKAZE PRI SVOJOM

Šifrový text:

PTRTF AXWOEAN DTZTK MTDFTIG HSNFX RGAQMT HKO LCGPGD

Substitučná šifra

Zmeníme kľúč:

Nový kľúč:

ABCDEFGHIJKLMNOPQRSTUVWXYZ

QWERTYUIODASPFGHJKLZXCVBNM

Šifrový text:

PTRTF AXWOEAN DTZTK MTDFTIG HSNFX RGAQMT HKO LCGPGD

Nový otvorený text:

MEDEN KUBICKY JETER ZEJNEHO PLYNU DOKAZE PRI SVOMOJ

spravili sme malú zmenu kľúča

dostali sme skoro taký istý text

Lámanie šifry

Schéma hill-climbingu:

- Vygenerujeme náhodný kľúč
- Dokola:
  - Vyskúšame všetky lokálne zmeny
  - Pre každú vyhodnotíme výsledok
  - Vyberieme najlepšiu

  Prestaneme, ak už žiadna nezlepší
- Ak sa nám ešte chce, tak odznova

Kompresia

Prefixový kód

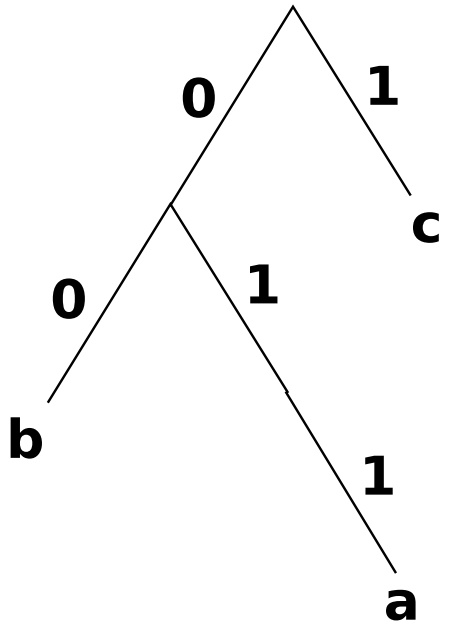
- postupnosť bitov = *kódové slovo*
- každému písmenu priradíme *kódové slovo*
- $\star$ žiadne kódové slovo nie je prefix iného
- príklad: ASCII kód (všetko má dĺžku 8)

Načo je nám $\star$?

Kompresia

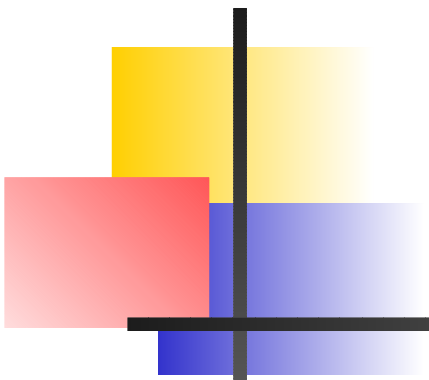
Prefixový kód = strom Príklad:

$a \to 011,\ b \to 00,\ c \to 1.$



Prefixový kód
$\leftrightarrow$ písmená sú listy
(čo ak navyše $d \to 01$?)

Je tento optimálny?
Je dobrý pre slovenčinu?



Kompresia

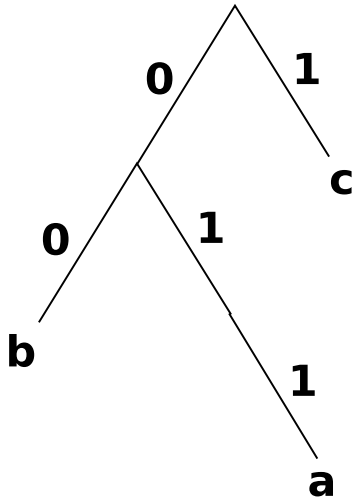
Optimálny prefixový kód

- žiadne zbytočné vnútorné vrcholy
- čím častejšie písmeno, tým kratší kód ($\Rightarrow$ šmejdy sú dole)
- čo ešte?

Kompresia

Optimálny prefixový kód

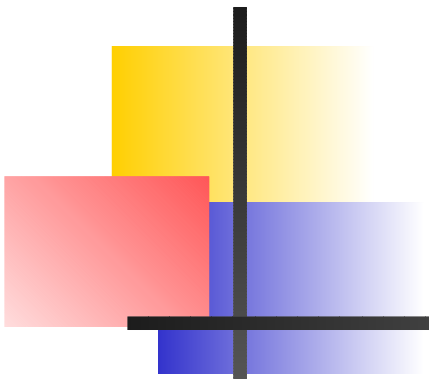
Každý vnútorný vrchol má stupeň 2:



(V tomto prípade a vie ísť hore.)

$\Rightarrow$ na spodnej úrovni sú aspoň 2 vrcholy

$\Rightarrow$ prehádžeme ich tak, aby 2 najväčšie šmejdy susedili



Kompresia

Huffmanov algoritmus

- vstup: pre každý znak pp výskytu
- priebeh: kým mám aspoň 2 znaky, opakuj:
  - nájdi dva najmenej časté
  - nahraď jedným novým so spoločnou pp
- implementácia: $O(n^2)$ pole, $O(n \log n)$ halda

Kompresia

Huffmanov algoritmus – vylepšenia

- využijeme bigramy (alebo viac :-)
- pp závisí od predchádzajúceho písmena
  $\Rightarrow$ nech závisí aj strom!
- SVK text bez diakritiky:
  - plain: 4.75 bitu / znak
  - 1D Huffman: 4.26 bitu / znak
  - 2D Huffman: 3.59 bitu / znak
  - 3D Huffman: 3.11 bitu / znak
  - 4D Huffman: 2.50 bitu / znak (ale obrovský strom)
  - **nekonverguje** k nule (ale rádovo k 1)!