MI–SPI První úloha

Tomáš Pšenička, Jan Groschaft

9. května 2020

Obsah

1 Zadání

- Z obou datových souborů načtěte texty k analýze. Pro každý text zvlášť odhadněte pravděpodobnosti znaků (symbolů včetně mezery), které se v textech vyskytují.
- Výsledné pravděpodobnosti graficky znázorněte. Pro každý text zvlášť spočtěte entropii odhadnutého rozdělení znaků.
- Nalezněte optimální instantní kód CC pro kódování znaků jednoho z textů.
- Pro každý text zvlášť spočtěte střední délku kódu CC a porovnejte ji s entropií rozdělení znaků.

2 Řešení

2.1 Pravděpodobnosti znaků

Pravděpodobnost jednotlivých znaků obou textů je zaznamenána v tabulkách 2 a 3. Pravděpodobnost jednotlivých znaků byla vypočítána jako poměr výskytu znaku k celkovému počtu všech znaků a je znázorněna v grafech 1 a 2 pomocí nástroje matplotlib.

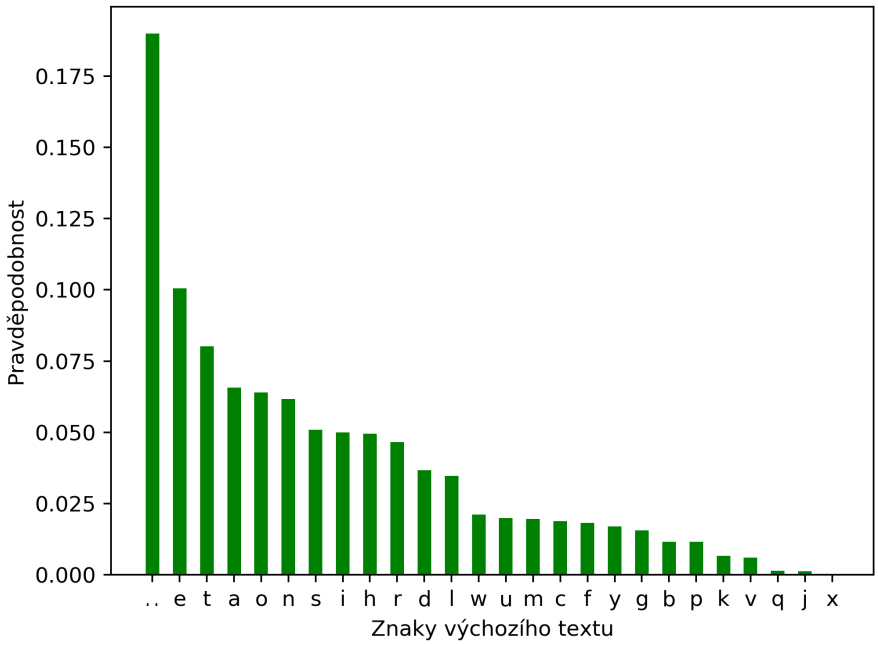
2.2 Entropie textů

Entropie obou textů byla vypočítána na základě zjištěných pravděpodobností následujícím vzorcem.

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

Text	Entropie
009.txt	4.060006550506617
011.txt	4.0639057100402844

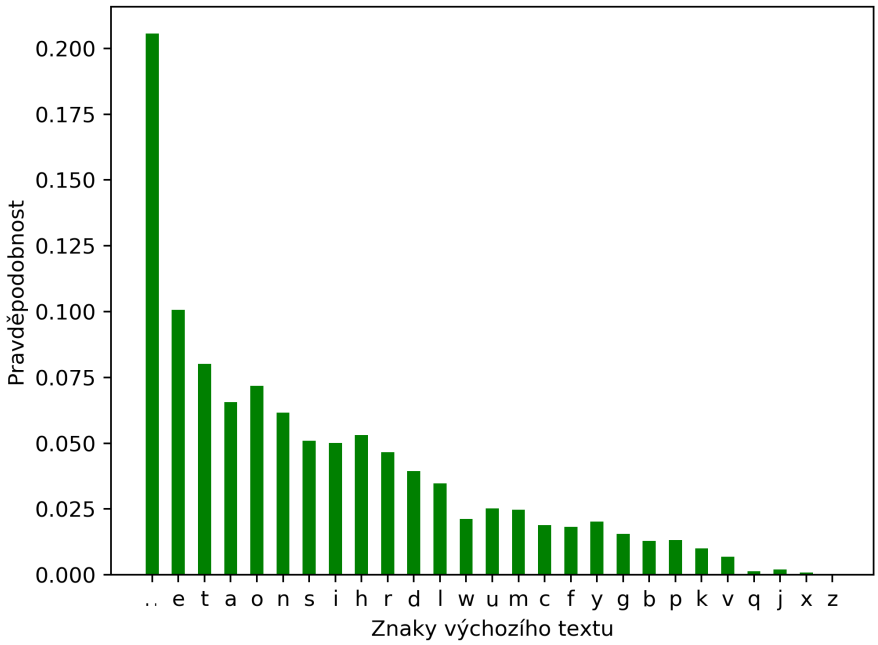
Tabulka 1: Entropie znaků obou textů



Obrázek 1: Grafické znázornění pravděpodobností znaků textu 009.txt

2.3 Optimální instantní kód

Pro nalezení optimálního kódu byl použit algoritmus sestrojení binárního Huffmanova kódu. Implementace tohoto algoritmu je provedena pomocí prioritní fronty a binárního stromu. Fronta je naplněna stavy, které obsahují jednotlivé znaky a je seřazena podle jejich pravděpodobnosti. Dva stavy s nejmenší pravděpodobností jsou vždy spojeny v jeden nový, který je zařazen do binárního stromu. Takto vytvořený stav obsahuje ukazatele na potomky, ze kterých byl sestaven a jejich sečtenou pravděpodobnost.
Tento postup se opakuje, dokud nezůstane poslední stav ve frontě. Výsledný kód je získán průchodem vzniklým binárním stromem až do listů. Cesta do listu určuje podobu kódového slova znaku, který se v něm nachází.
Takto nalezené optimální kódy pro oba texty jsou zobrazeny v tabulkách 2 a 3.



Obrázek 2: Grafické znázornění pravděpodobností znaků textu 011.txt

2.4 Střední délka kódu

Střední délka optimálního kódu sestaveného pro text 009.txt je pro oba texty vypočítána pomocí následujícího vzorce.

$$L(C) = \sum_{x \in X} l(x)p(x)$$

Znak	Pravděpodobnost	Kód
”␣”	0.1898870360408822	00
”e”	0.10059171597633136	010
”t”	0.08015061861215707	1101
”a”	0.0656266810112964	1010
”o”	0.06401291016675632	1001
”n”	0.06168190783575399	1000
”s”	0.05092343553882015	0110
”i”	0.05002689618074233	11111
”h”	0.04948897256589564	11110
”r”	0.046620046620046623	11101
”d”	0.036758113681190606	11000
”l”	0.0347857270934194	10110
”w”	0.021158328850636544	111001
”u”	0.019903173749327596	111000
”m”	0.019544558006096467	110011
”c”	0.018827326519634213	110010
”f”	0.01828940290478752	101111
”y”	0.017034247803478574	101110
”g”	0.015599784830554062	011110
”b”	0.011655011655011656	011100
”p”	0.011655011655011656	011101
”k”	0.006634391249775865	0111110
”v”	0.006096467634929173	01111111
”q”	0.0014344629729245114	011111100
”j”	0.0012551551013089475	0111111011
”x”	0.0003586161574323112785	0111111010

Tabulka 2: Pravděpodobnosti znaků a Huffmanův kód sada 009.txt.

Znak	Pravděpodobnost	Kód
”␣”	0.20561816103217376	01
”e”	0.08835538134901193	1110
”t”	0.07724971419238935	1100
”o”	0.07186019924873428	1011
”a”	0.05879470847623714	1001
”h”	0.05307855626326964	0011
”n”	0.0514453699167075	0010
”i”	0.049158909031520495	0000
”s”	0.04883227176220807	11111
”r”	0.04246284501061571	11110
”d”	0.03935979095214764	11010
”l”	0.03184713375796178	10100
”u”	0.025314388371713212	00011
”m”	0.024661113833088357	00010
”y”	0.02025151069737057	110110
”w”	0.019434917524089497	101011
”f”	0.01649518210027764	101010
”c”	0.015188633023027927	100011
”g”	0.014045402580434428	100010
”p”	0.013228809407153356	100001
”b”	0.012902172137840928	100000
”k”	0.010125755348685286	1101110
”v”	0.006859382655560999	11011111
”j”	0.0019598236158745713	110111101
”x”	0.0008165931732810714	1101111001
”q”	0.0004899559039686428	11011110001
”z”	0.00016331863465621427	11011110000

Tabulka 3: Pravděpodobnosti znaků a Huffmanův kód sada 011.txt.

Text	Entropie	Délka CC kódu pro 009.txt	Délka CC kódu pro 011.txt
009.txt	4.060006550506617	4.105791644253182	4.119956966110813
011.txt	4.0639057100402844	4.1156295933366	4.101420872121509

Tabulka 4: Délka Huffmanova kódu sestaveného pro oba texty

3 Závěr

Nejpravděpodobnějším znakem obou zkoumaných textů byla mezera. Jako další dva nejvíce pravděpodobné znaky vychází písmena „e“ a „t“. To odpovídá standardní frekvenci těchto znaků pro anglický text.
Entropie byla pro oba texty téměř shodná, po zaokrouhlení rovna 4,1.
Optimální kód byl nalezen pomocí Huffmanova algoritmu pro oba texty. Kódová slova pro nejvíce frekventované znaky jsou dle očekávaní nejkratší. Optimální kód pro druhý text nebylo dle zadání potřeba konstruovat, jeho podobu uvádíme pouze pro porovnání.
Pro optimální kód platí dle předpokladů nerovnost:

$$H(X) \leq L(CC) < H(X) + 1$$