MI–SPI Druhá úloha

Tomáš Pšenička, Jan Groschaft

3. května 2020

Obsah

1 Zadání

- Z obou datových souborů načtěte texty k analýze. Pro každý text zvlášť odhadněte základní charakteristiky délek slov, tj. střední hodnotu a rozptyl. Graficky znázorněte rozdělení délek slov.
- Pro každý text zvlášť odhadněte pravděpodobnosti písmen (symbolů mimo mezery), které se v textech vyskytují. Výsledné pravděpodobnosti graficky znázorněte.
- Na hladině významnosti 5% otestujte hypotézu, že rozdělení délek slov nezávisí na tom, o který jde text. Určete také p-hodnotu testu.
- Na hladině významnosti 5% otestujte hypotézu, že se střední délky slov v obou textech rovnají. Určete také p-hodnotu testu.
- Na hladině významnosti 5% otestujte hypotézu, že rozdělení písmen nezávisí na tom, o který jde text. Určete také p-hodnotu testu.

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.

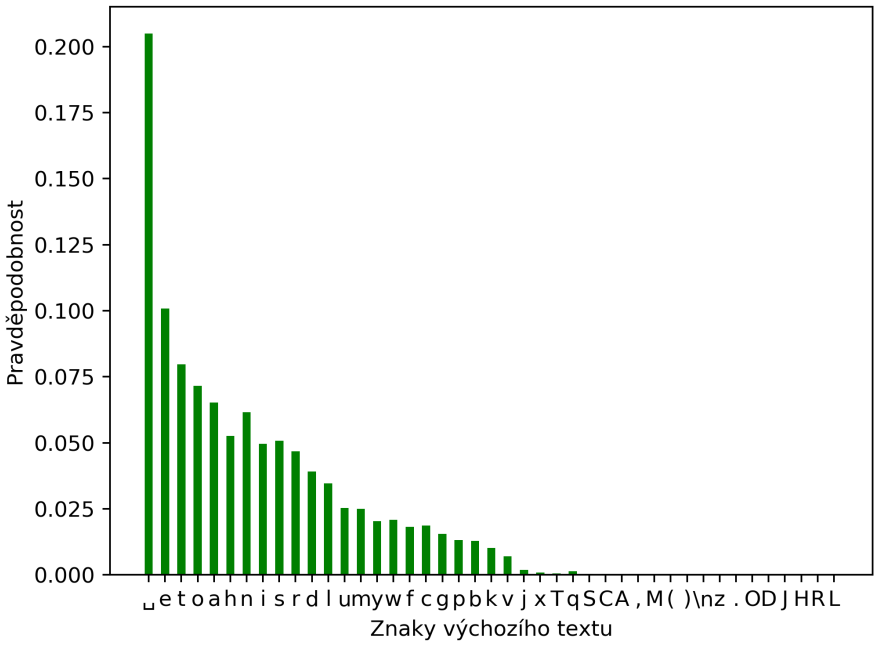
2.2 Entropie textů

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	4.087103237335235
011.txt	4.084709593540786

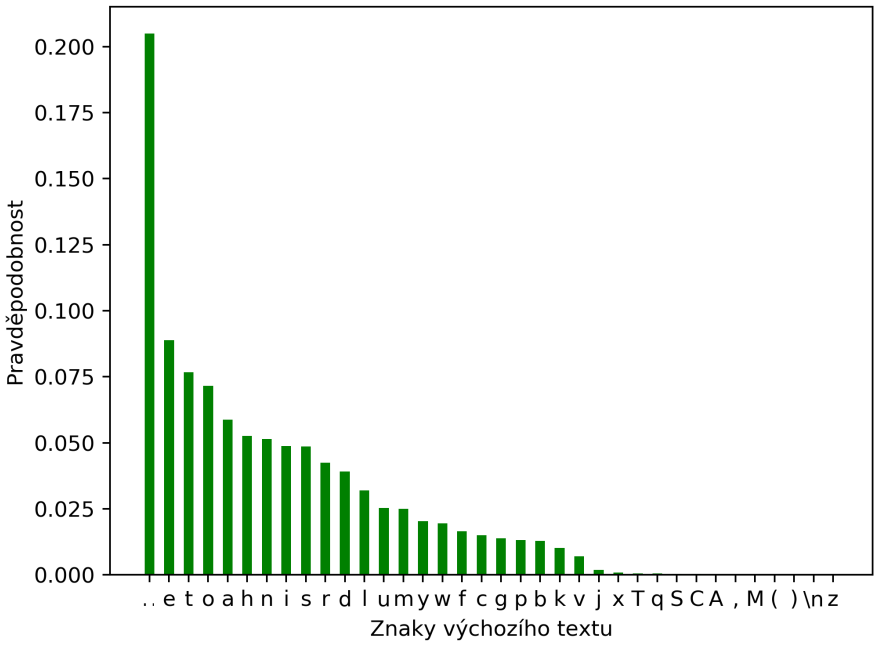
Tabulka 1: Entropie znaků obou textů



Obrázek 1: Grafické znázornění pravděpodobností znaků textu 009.txt

2.3 Optimální instantní kód

Pro nalezení optimálního kódu byl použit algoritmus sestrojení binárního Hu□manova 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
”␣”	0.18962464589235128	00
”e”	0.10074362606232294	010
”t”	0.07967422096317281	1101
”a”	0.06515580736543909	1010
”o”	0.06373937677053824	1001
”n”	0.06161473087818697	1000
”s”	0.05081444759206799	0110
”i”	0.049575070821529746	11111
”h”	0.049043909348441925	11110
”r”	0.046742209631728045	11101
”d”	0.03665014164305949	11000
”l”	0.034702549575070823	10110
”w”	0.020892351274787536	111001
”u”	0.019830028328611898	111000
”m”	0.01929886685552408	110011
”c”	0.018590651558073653	110010
”f”	0.01823654390934844	101110
”y”	0.017351274787535412	011111
”g”	0.015580736543909348	011110
”b”	0.011862606232294617	011100
”p”	0.011508498583569405	1011111
”k”	0.006728045325779037	1011110
”v”	0.006196883852691218	0111011
”q”	0.00141643059490085	011101001
”j”	0.0012393767705382436	011101000
”S”	0.0003541076487252125	011101011110
”.”	0.0003541076487252125	011101011111
”x”	0.0003541076487252125	01110101000
”T”	0.00017705382436260624	011101010010
”C”	0.00017705382436260624	011101010011
”O”	0.00017705382436260624	011101010100
”D”	0.00017705382436260624	011101010101
”J”	0.00017705382436260624	011101010110
”A”	0.00017705382436260624	011101010111
”M”	0.00017705382436260624	011101011000
”H”	0.00017705382436260624	011101011001
”,”	0.00017705382436260624	011101011010
”R”	0.00017705382436260624	011101011011
”L”	0.00017705382436260624	011101011100
”\n”	0.00017705382436260624	011101011101

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

Znak	Pravděpodobnost	Kód
”␣”	0.20490876796383012	01
”e”	0.08880994671403197	1110
”t”	0.07669949943484579	1100
”o”	0.07153237526239302	1011
”a”	0.05877603746165025	1001
”h”	0.05264007750686259	0011
”n”	0.051348296463749395	0010
”i”	0.04876473437752301	0000
”s”	0.048603261747133863	11111
”r”	0.042467301792346195	11011
”d”	0.03907637655417407	11010
”l”	0.03197158081705151	10100
”u”	0.025351202971096398	00011
”m”	0.0250282577103181	00010
”y”	0.020345551429032778	111100
”w”	0.019538188277087035	101011
”f”	0.0164702082996932	101010
”c”	0.01501695462619086	100011
”g”	0.013886646213466818	100010
”p”	0.013240755691910222	100001
”b”	0.012917810431131924	100000
”k”	0.010172775714516389	1111010
”v”	0.006943323106733409	11110111
”j”	0.0019376715646697885	1111011011
”x”	0.0008073631519457452	11110110101
”T”	0.00048441789116744714	11110110000
”q”	0.00048441789116744714	11110110001
”S”	0.0003229452607782981	111101100101
”C”	0.0003229452607782981	111101100110
”A”	0.00016147263038914905	1111011001110
”,”	0.00016147263038914905	1111011001111
”M”	0.00016147263038914905	1111011010000
”(”	0.00016147263038914905	1111011010001
”)”	0.00016147263038914905	1111011010010
”\n”	0.00016147263038914905	1111011010011
”z”	0.00016147263038914905	111101100100

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	4.087103237335235	4.131728045325779	4.12304214435653
011.txt	4.084709593540786	4.134506701114161	4.127832861189802

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:

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