MI–SPI Druhá úloha

Tomáš Pšenička, Jan Groschaft

10. 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 Charakteristika délek slov

Pro odhad charakteristik délek slov jsme použili výběrový průměr a výběrový rozptyl:

$$\bar{X}_n = \frac{1}{n} \sum_i X_i$$

$$s_n^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2$$

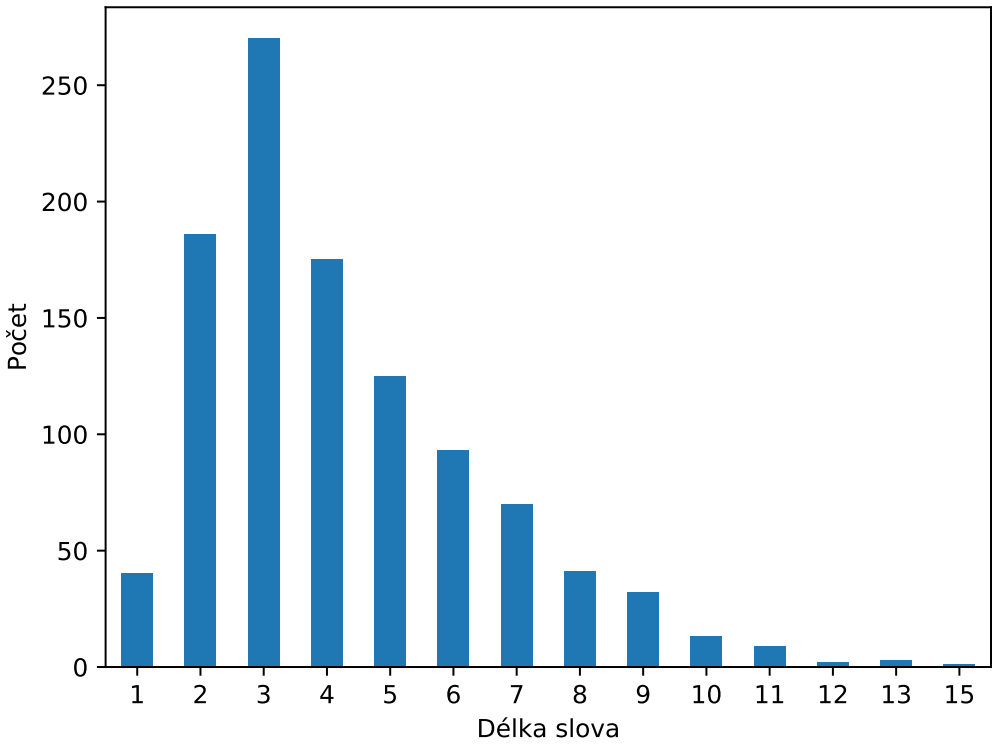
Výsledky sledujeme v tabulce 1 a na obrázcích 1 a 2.

Text	$\bar{X}_n$	s_n^2
009.txt	4.2623	4.9860
011.txt	3.8603	4.5360

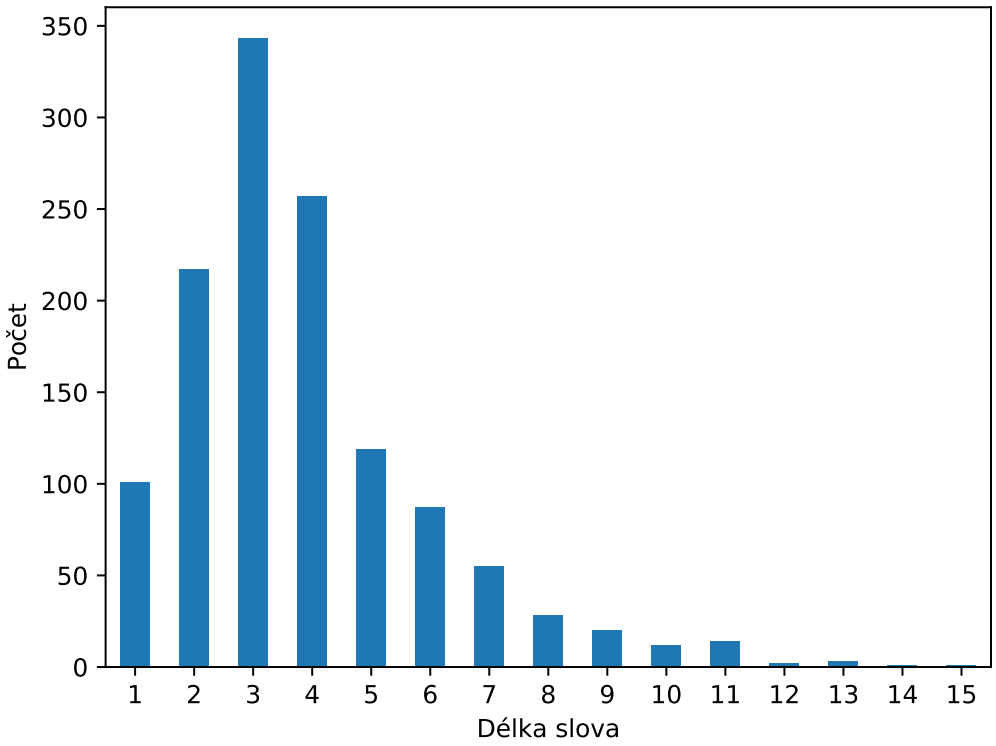
Tabulka 1: Charakteristika délek slov

2.2 Pravděpodobnosti znaků

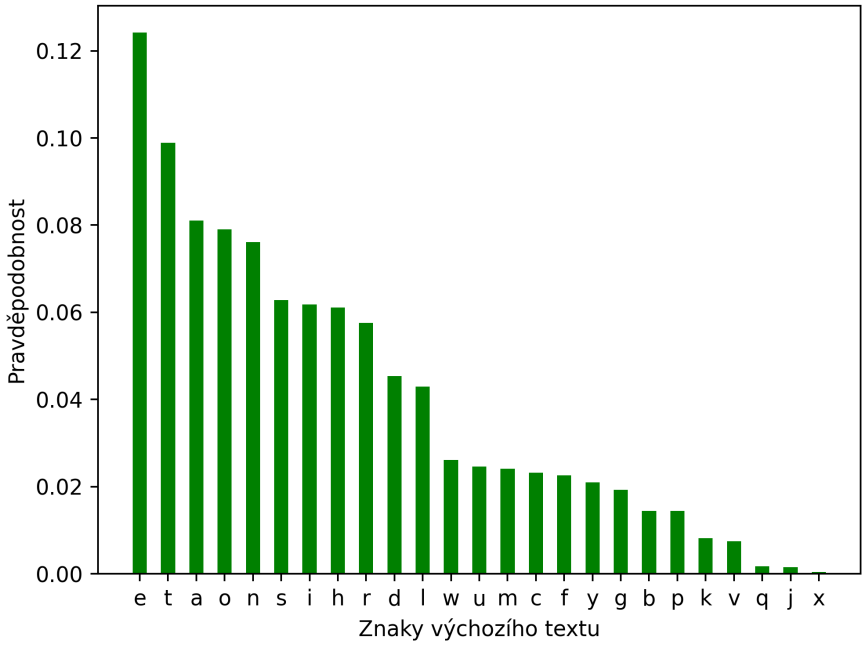
Pravděpodobnost jednotlivých znaků byla vypočítána jako poměr výskytu znaku k celkovému počtu všech znaků a je znázorněna v grafech 3 a 4 pomocí nástroje matplotlib.



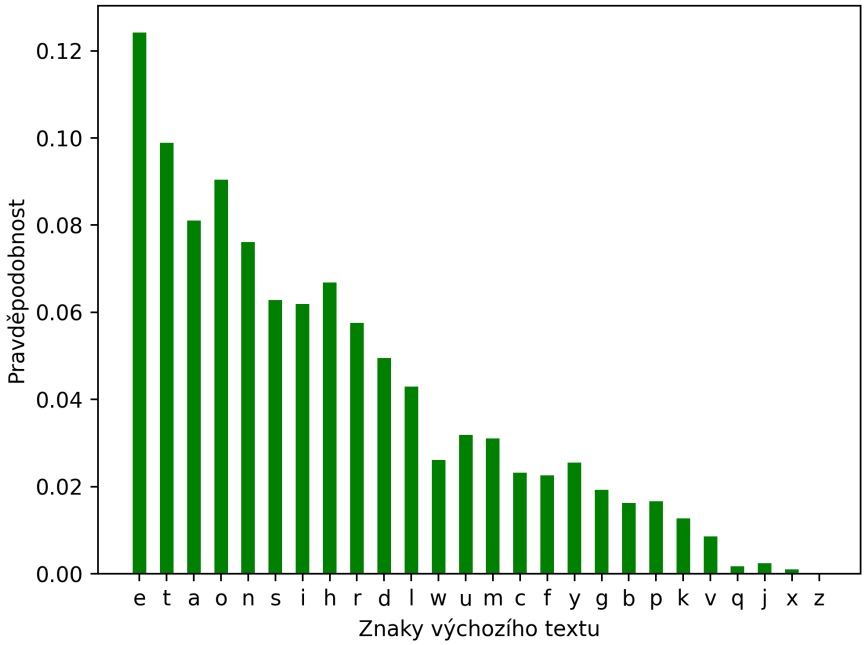
Obrázek 1: Rozdělení délek slov v souboru 009.txt



Obrázek 2: Rozdělení délek slov v souboru 011.txt



Obrázek 3: Grafické znázornění pravděpodobností znaků textu 009.txt



Obrázek 4: Grafické znázornění pravděpodobností znaků textu 011.txt

3 Závěr