**Domácí úkol č. 2**

Karel Vrabec

**Zadání**

Obsah obrázku text

Popis byl vytvořen automaticky

**Řešení**

Testované soubory: **005.txt** (jako **text č. 1**), **017.txt** (jako **text č. 2**)

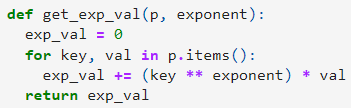
Programovací jazyk: **Python**

Kódy k dispozici v přiloženém archivu **source.zip** (soubor **assignment.ipynb**)

**1.** Načetl jsem texty z obou datových souborů. Pro každý z nich jsem vypočítal **četnosti** jednotlivých délek (tzn. celkový počet slov pro každou délku). Následně jsem tyto četnosti vydělil celkovým počtem slov v daném textu, čímž jsem získal **pravděpodobnosti** těchto délek. Tyto odhadnuté pravděpodobnosti jsem použil při výpočtu střední hodnoty a rozptylu.

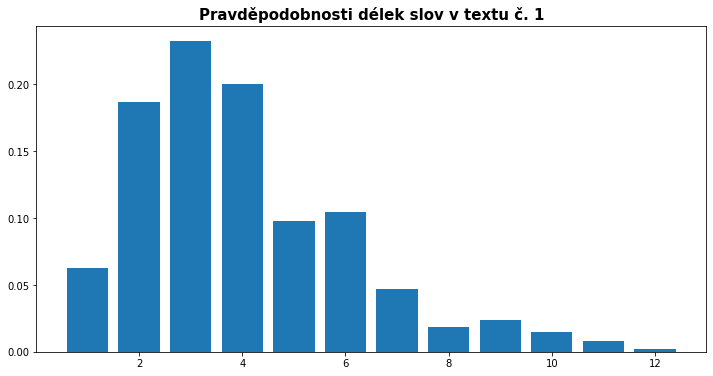
|  |  |  |
| --- | --- | --- |
|  |  |  |
| Text č. 1 |  |  |
| Text č. 2 |  |  |

Příslušná ukázka kódu (výpočet střední hodnoty a rozptylu):



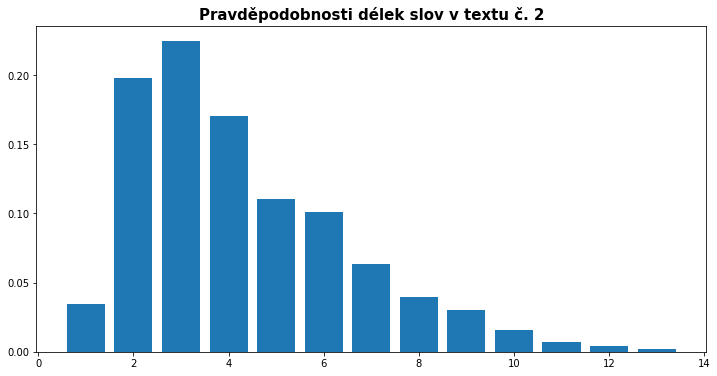


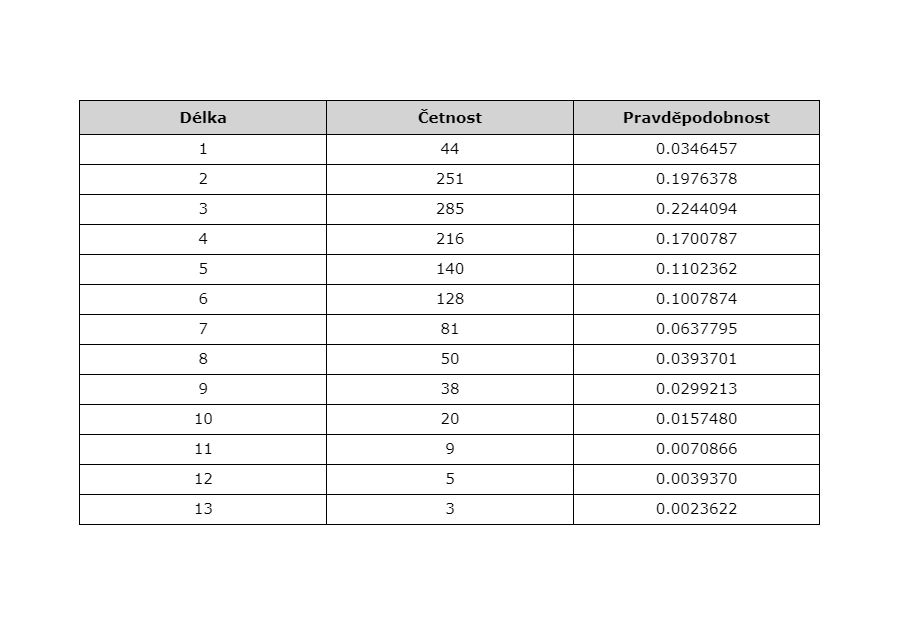
Odhadnuté pravděpodobnosti jsem také graficky znázornil v následujících grafech. Slova o délce **3** jsou v obou úryvcích nejvíce používanými slovy. Další délky vykazují klesající charakter.



Obsah obrázku stůl

Popis byl vytvořen automaticky



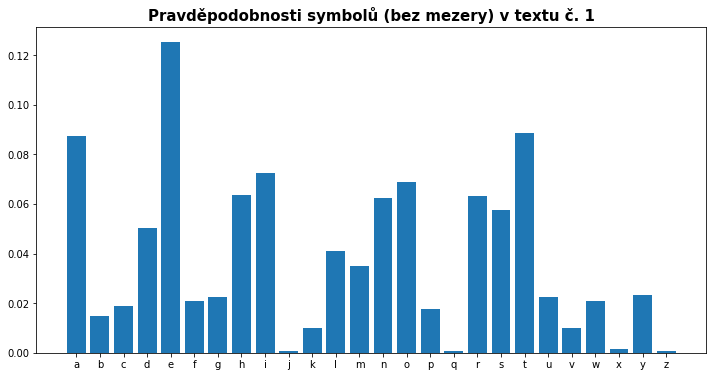


Příslušná okomentovaná ukázka kódu (načítání textů, výpočet četností a pravděpodobností):

Obsah obrázku text

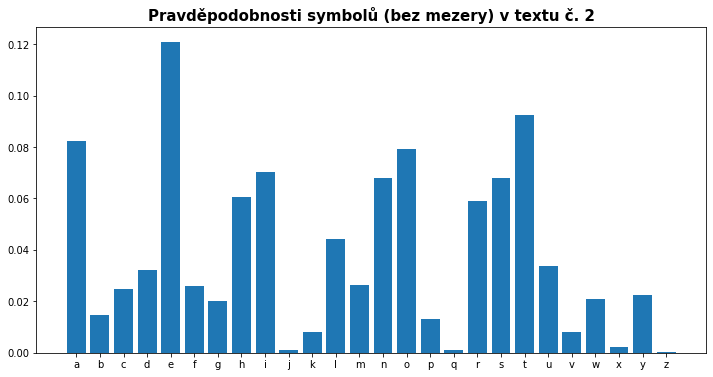
Popis byl vytvořen automaticky

**2.** Pro každý z textů jsem vypočítal **četnosti** jednotlivých znaků (bez mezery). Následně jsem tyto četnosti vydělil celkovým počtem znaků v daném textu (bez mezer), čímž jsem získal **pravděpodobnosti** těchto znaků. Odhadnuté pravděpodobnosti jsem graficky znázornil v následujících grafech. Písmena **j**, **q**, **x** a **z** jsou v obou úryvcích nejméně používanými znaky a písmeno **e** naopak tím nejpoužívanějším.



Obsah obrázku stůl

Popis byl vytvořen automaticky



Obsah obrázku stůl

Popis byl vytvořen automaticky

**3.** Provedl jsem **test nezávislosti v kontingenční tabulce**, kdy testuji **nulovou** hypotézu (existence nezávislosti rozdělení délek slov) vůči **alternativní** (neexistence nezávislosti rozdělení délek slov) na hladině významnosti .

Obsah obrázku text

Popis byl vytvořen automaticky

Pro tento test jsem nejdříve vytvořil následující kontingenční tabulku. Četnosti slov délek **11, 12, 13** jsem sloučil, aby jejich **teoretické četnosti** dosahovaly hodnoty **alespoň 5**.

Obsah obrázku stůl

Popis byl vytvořen automaticky

Na základě této tabulky jsem vypočetl **testovou statistiku** () a **kritickou hodnotu** (). Vzhledem k tomu, že platí (tzn. že se nacházíme v **kritickém oboru**), **zamítám** nulovou hypotézu o nezávislosti rozdělení délek slov ve prospěch hypotézy alternativní. **P-hodnotou** je .

**4.** Provedl jsem **dvouvýběrový t-test za předpokladu různých rozptylů**, kde testuji **nulovou** hypotézu (rovnost středních délek slov) vůči **alternativní** (nerovnost středních délek slov) na hladině významnosti .

Obsah obrázku stůl

Popis byl vytvořen automaticky

Na základě uvedených vztahů jsem vypočítal **testovou statistiku** () a **kritickou hodnotu** (). Vzhledem k tomu, že platí   
(tzn. že se nacházíme v **kritickém oboru**), **zamítám** nulovou hypotézu o rovnosti středních délek slov ve prospěch hypotézy alternativní. **P-hodnotou** je .

Příslušná okomentovaná ukázka kódu (výpočet testové statistiky a kritické hodnoty):

Obsah obrázku text

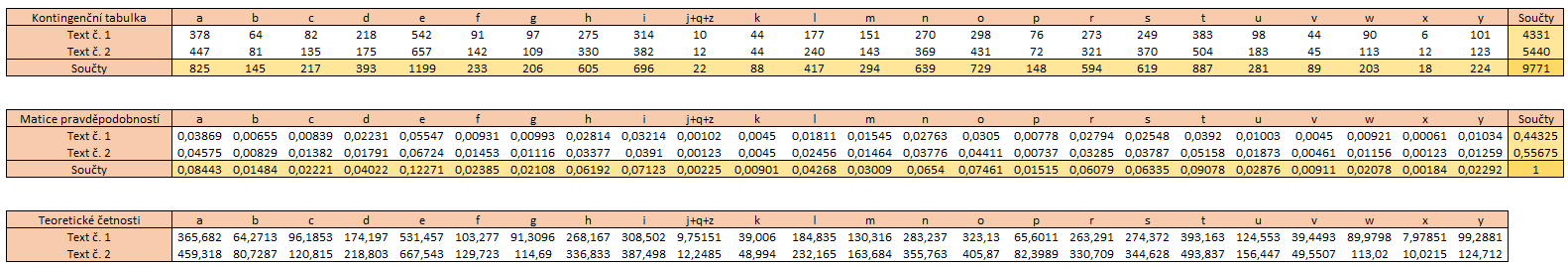
Popis byl vytvořen automaticky

**5.** Provedl jsem **test nezávislosti v kontingenční tabulce**, kde testuji **nulovou** hypotézu (existence nezávislosti rozdělení písmen) vůči **alternativní** (neexistence nezávislosti rozdělení písmen) na hladině významnosti .

Obsah obrázku text

Popis byl vytvořen automaticky

Pro tento test jsem nejdříve vytvořil následující kontingenční tabulku. Četnosti písmen **j**, **q** a **z** jsem sloučil, aby jejich **teoretické četnosti** dosahovaly hodnoty **alespoň 5**.



Na základě této tabulky jsem vypočetl **testovou statistiku** () a **kritickou hodnotu** (). Vzhledem k tomu, že platí (tzn. že se nacházíme v **kritickém oboru**), **zamítám** nulovou hypotézu o nezávislosti rozdělení písmen ve prospěch hypotézy alternativní. **P-hodnotou** je .