FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

Semestrální práce z předmětu
Hlasové a dialogové systémy

Fonetická transkripce textu

Autoři:
Daniel Cífka
A23N0100P
dcifka20@students.zcu.cz

Stanislav Kafara
A24N0088P
skafara@students.zcu.cz

Obsah

Kapitola 1

Zadání

Vytvořte skript v programovacím jazyce Python, který provede fonetickou transkripci textového souboru. Pro definici fonetických transkripčních pravidel využijte [1] (podklady jsou na CourseWARE). Soustřeďte se na fonetickou transkripci spisovné české výslovnosti. Alternativně je možné (bez možnosti získat bonusové body) rovněž využít volně dostupné nástroje phonemizer (s backendem speak-ng) a gruut (podrobnosti budou uvedeny na cvičení). Tyto nástroje pracují s fonetickou abecedou IPA, výstup tedy bude nutné převést do abecedy EPA (viz dále).

Vstup

Textový soubor s ortografickou transkripcí `vety_HDS.ortho.txt`, jehož každý řádek obsahuje samostatnou větu nebo souvětí.

Ukázka vstupu

> okolo se prohánějí skejťáci, zájem o něj projevují také cizinci.

Výstup

Textový soubor s fonetickou transkripcí `vety_HDS.phntrn.txt`, jehož každý řádek obsahuje foneticky přepsanou větu. Pořadí foneticky přepsaných a původních ortograficky zapsaných vět musí být stejné. K fonetickému zápisu použijte dodanou fonetickou abecedu EPA.

Ukázka výstupu

$!okolo se prohAJejI skejTAci, # zAjem !o Jej projevujI také cizinci. $

Poznámky

Symbol „$“ označuje „delší“ pauzu na začátku a konci věty. Symbol „#“ označuje „kratší“ pauzu uvnitř vět. Krátké pauzy vkládejte na místa, kde se v původním textu vyskytovala interpunkce. Pauzy mají charakter neznělých zvuků.

Součásti zadání

- Fonetická transkripční pravidla [1] (jsou součástí podkladů na CourseWARE)
- `vety_HDS.ortho.txt` – vstupní textový soubor s ortografickou transkripcí
- `ukazka_HDS.ortho.txt` – ukázkový textový soubor s ortografickou transkripcí
- `ukazka_HDS.phntrn.txt` – ukázkový výstupní textový soubor s fonetickou transkripcí
- `phonetic_alphabet_cz.pdf` – fonetická abeceda (značená jako EPA)

Zájemci o bonusové body odevzdají pomocí Google Classroom své výstupní fonetické transkripce souboru `vety_HDS.ortho.txt` pojmenované jako `vety_HDS.phntrn.txt` a pythonovský skript provádějící transkripci.

[1] PSUTKA, J., MÜLLER, L., MATOUŠEK, J., RADOVÁ, V. Mluvíme s počítačem česky. Kapitola 2: Vytváření a vlastnosti mluvené řeči. Academia, Praha, 2006.

Kapitola 2

Popis implementace

2.1 Načtení dat `data_loader()`

Funkce `data_loader()` načítá jednotlivé řádky ze vstupního souboru a připravuje je k následnému zpracování.

2.2 Předzpracování `preprocesing()`

Funkce `preprocesing()` zajišťuje základní úpravy vstupního textu:

- odstranění nepotřebných značek,
- nahrazení interpunkčních znamének požadovanými značkami,
- vrácení seznamu přepsaných vět pro další zpracování.

2.3 Fonetická transkripce `check_rules()`

Tato část skriptu aplikuje fonetická pravidla pomocí podmínek `if` a `elif` pro přepis vět do fonetické podoby textu.

- Výstupem funkce `check_rules()` je seznam, jehož délka odpovídá počtu znaků ve větě.
- V seznamu jsou na pozicích odpovídajících přepsaným znakům uloženy nové prvky, zatímco ostatní pozice obsahují hodnotu `None`.

2.4 Výstup a převod `final_preprocesing_and_write()`

Funkce `final_preprocesing_and_write()` zajišťuje:

- převod znaků `None` na znaky z původní věty,
- spojení prvků do výsledného řetězce,
- převod znaků do abecedy EPA,
- převod unikátních slov na jejich fonetickou podobu,
- zápis výsledných vět do výstupního souboru.

Kapitola 3

Nevyužitá fonetická pravidla

Níže je uveden seznam fonetických pravidel, která nebyla ve skriptu implementována:

```
- ZPK   -> ¬ZPK   /_<+JK>,
- ZPK   -> ZPK    /_<+JK>,
- ZPK1  -> ZPK1   /<"JPZ">_<-ZPK2,-JK>,
- t     -> ť      /_<ň>,
- t     -> t      /_<ň>,
- d     -> ď      /_<ň>,
- d     -> d      /_<ň>,
- n     -> ň      /_<ť,ď>,
- n     -> n      /_<ť,ď>,
- t-š   -> tš,
- t|š   -> tš,
- tš    -> č,
- tš    -> tš,
- d-z   -> dz,
- d|z   -> dz,
- dz    -> dz,
- d-ž   -> dž,
- d|ž   -> dž,
- dž    -> dž,
- zští  -> šťí    /_|,
- zští  -> sšťí   /_|,
- žští  -> šťí    /_|,
- žští  -> ššťí   /_|,
- ť     -> t      /š_k,
- ť     -> ť      /š_k,
```

```
- d     -> d      /z_<"n,ň">,
- d     -> i      /z_<"n,ň">,
- vz    -> z      /<|,-,>_<-b,-p>,
- fs    -> s      /<|,-,>_<-b,-p>,
- Výslovnost foneticky stejných souhlásek (všechna pravidla)
```

Kapitola 4

Uživatelská příručka

Spuštění skriptu

Skript se spouští pomocí příkazové řádky:

```
python main.py -inputfile <in> -outputfile <out> [-split]
```

- `<in>` – cesta k vstupnímu souboru s textem
- `<out>` – cesta k výstupnímu souboru s fonetickou transkripcí textu
- `-split` - přepínač indikující skriptu nutnost odděleného zpracování textu a identifikátoru textu ve vstupním souboru

Příklad spuštění

Pokud vstupní soubor obsahuje jen text, např.:

```
za svůj gól však musím poděkovat především kovalevovi.
ani po zhroucení však svoboda nebude na cele sám.
...
```

pak skript pro vstupní soubor `vety_HDS.ortho.txt` s požadovaným výstupním souborem `vety_HDS.phntrn.txt` spustíme následovně:

```
python main.py -inputfile vety_HDS.ortho.txt
-outputfile vety_HDS.phntrn.txt
```

Výsledný soubor obsahující fonetickou transkripci vypadá následovně:

```
$ za zvUj gOl fSak musIm poDekovat pRedefSIm kovalevovi. $
$ aJi po zhrouceJI fSak zvoboda nebude na cele sAm. $
```

Pokud vstupní soubor obsahuje text a zároveň identifikátor textu, např.:

```
Sentence00001|tím předsedkyně senátu okresního soudu ...
Sentence00002|nadstandardní zdravotní péče zmíněným n...
...
```

pak skript pro vstupní soubor `sentences_phnfriendly.csv` s požadovaným výstupním souborem `train.ph-redu.epa.csv` spustíme následovně:

```
python main.py -inputfile sentences_phnfriendly.csv
-outputfile train.ph-redu.epa.csv -split
```

Výsledný soubor obsahující fonetickou transkripci vypadá následovně:

```
Sentence00001|$ TIm pRetsetkiJe senAtu okresJIho soudu ... $
Sentence00002|$ natstandardJI zdravotJI pECe zmIJenIm n... $
```