Zdroj dát

Ako zdroj textových dát bola zvolená slovenská Wikipédia. Tento zdroj poskytuje veľké množstvo verejne dostupného textu v slovenskom jazyku, ktorý je vhodný na štatistickú analýzu jazyka a tvorbu slovníkov. Na extrakciu čistého textu z XML dumpu bol použitý nástroj WikiExtractor.

```
python -m wikiextractor.WikiExtractor skwiki-latest-pages-articles.xml.bz2 -o wiki_text --no-templates
```

Spracovanie textu a filtrácia

Extrahovaný text bol spracovaný pomocou vlastných Python skriptov. Počas spracovania boli aplikované nasledujúce filtre:

1. konverzia slov na malé písmená
2. odstránenie neabecedných znakov
3. ponechanie iba slov so slovenskými diakritickými znakmi
4. odstránenie technických výrazov (napr. wiki, html, xml)

```
cat wiki_text/*/* | python3 scripts/count_frequencies.py > intermediate/frequencies.txt
```

Normalizácia frekvencií

Surové frekvencie boli následne normalizované pomocou logaritmickej funkcie, aby sa zabránilo dominancii veľmi častých slov. Maximálna hodnota frekvencie bola obmedzená na 255, čo zodpovedá rozsahu používanému v LatinIME.

```
python3 normalize_frequencies.py < ../intermediate/frequencies.txt > ../intermediate/normalized.txt

zip sk_wordlist.combined.zip sk_wordlist.combined
```

Výsledok

Výsledkom projektu je kvalitný frekvenčný slovník slovenského jazyka, ktorý je pripravený na ďalšie spracovanie pomocou nástroja dicttool_aosp v prostredí AOSP. Projekt preukázal, že príprava jazykového slovníka je realizovateľná mimo systému Android, avšak jeho integrácia a testovanie si vyžaduje kompletné zostavenie systému GrapheneOS.