Model Creator

Aplikace slouží k vytvoření jazykového modelu z trénovacího korpusu. Využívá část ngram-count projektu SRILM, která slouží k vytváření modelů v rámci projektu SRILM. Znaky trénovacího korpusu rozdělí mezerami, tak jak to vyžaduje projekt SRILM, aby pracoval se znaky. Původní mezery nahradí za podtržítko „_“, proto by se v trénovacím korpusu nemělo podtržítko objevovat.

Povinné parametry

Parametr	Popis
input	Soubor s texty daného jazyka (trénovací korpus).
output	Název výstupního souboru (model).
vocabulary	Soubor slovníku, kde každý znak je na jednom řádku. Soubor musí být v kódování UTF-8 bez BOM.

Volitelné parametry

Parametr	Popis	Defaultní hodnota
order	Stupeň modelu, který se má vytvořit.	5
encoding	Kódování vstupního souboru.	UTF-8
discounting	Použitá vyhlazovací technika. Hodnoty tohoto paremetru jsou shodné jako ve SRILM.	-wbdiscount -interpolate

Příklad

ModelCreator input Czech.txt output Czech.lm vocabulary vocabulary.txt order 3 encoding windows-1250