**Nikud Remover How To Info**

Author: Benjamin Katz

Phone: +12017448809(Whatsapp)

Email: benjykatz12@gmail.com

Written: Summer 2022

Please reach out with any questions

**Running the Software (NR.py):**

This software was written to turn a .txt file with nikud(חסר) into a .txt file without nikud(מלא) this can be done by running *python NR.py <nikudTextFile>* and then going into the NikudRemover directory and looking at the NikudDeleted.txt file. This file should be manually reviewed to ensure that the algorithm is producing desirable results. This is really the only necessity for running the program assuming that all directories are correctly installed the Milon folder is proprietary to Bar Ilan so it is not on the git, ask Prof. Schler for a copy and place it directly into the NikudRemover folder.

**What to do with undesirable results:**

This is likely due to bad chaser-maleh dictionary inputs. The dictionaries can be manually reviewed and corrected; however, this is a large task and may not be worth it. More likely, it would be better to simply delete dictionaries that are based on bad inputs and replace them with dictionaries based on verified ‘Rosetta Stones’ of nikud and non nikud. New dictionaries can be created using the AddToDict.py program.

**AddToDict.py**

The AddToDict program is run by *python AddToDict.py <nikudText.txt> <nonNikudText.txt> <nameOfText>* and outputs a csv file into the ChaserToMalehDicts folder called chaserToMalehNameOfText.csv. The two files must be of the same text, small girsa and formatting changes are acceptable and delt with by the program, but large omissions of paragraphs and chapters could cause the program to output wrong results. This program could take a while to add all the words into the dictionary (I found it to take around an hour when adding close to the entirety of the Mishnah Torah on my personal machine). This likely can be optimized but because of the infrequency of adding to the dictionary it is not necessary. Important note: this program utilizes wdiff which is a linux utility program therefore this program will only run on a linux machine and wdiff must be installed beforehand. If you have pip installed, you can call *pip install wdiff*

**Incorporating the new dictionary:**

Any new dictionary must be manually added into NR.py program in lines 92-109 by creating a new pandas data frame that reads the csv and converts it to a dictionary and the ‘master’ dictionary needs to be updated to include this new dictionary.

**Data and fundamental Issues:**

The data used to build the original dictionaries come from machon mamre, wikitext, and the daf yomi portal. These were chosen because they have texts both with and without nikud. Some of the decisions that were made in the way of adding and removing nikud are not necessarily correct and the program is heavily reliant on those decisions. Additionally, there are different conventions as to how to convert chaser to maleh, depending on time period and authorship. This program makes no such distinctions.

**Files and directories:**

The root folder of the project is called NikudRemover, this folder contains the python programs along with input and output files. NikudDeleted.txt is the output file of NR.py, similarities.txt is a utility file that is utilized by AddToDict.py and never needs to be accessed by the end user. The Milon folder contains the frequency dictionaries which the program utilizes in the case that there is no Chaser-Maleh dictionary hit. The chaser-maleh dictionaries can be found in ChaserToMalehDicts folder as csv files. Pickle files are there to avoid unnecessary processing with every run of the program. All other files are either training or test data and not entirely necessary.