**Emoji, Emoticon, Abbreviations Extraction Tool Manual**

*Last Revised: 24 Oct 2022*

**Introduction**

The aim of this file is to extract emojis, emoticons and abbreviations from a certain data source. It aims to support the Diagnostic Tool by providing the above extracted features with meaningful interpretations so that information loss is minimized.

**Main File**

*Preprocessing.ipynb*

This is a Google Colaboratory file run on Google Drive service. The advantage of doing so is the drastic increase in speed that the free GPU provided by Google services as compared to reserving computing power on local machines for the task.

**Input Files**

1. Labeled data set

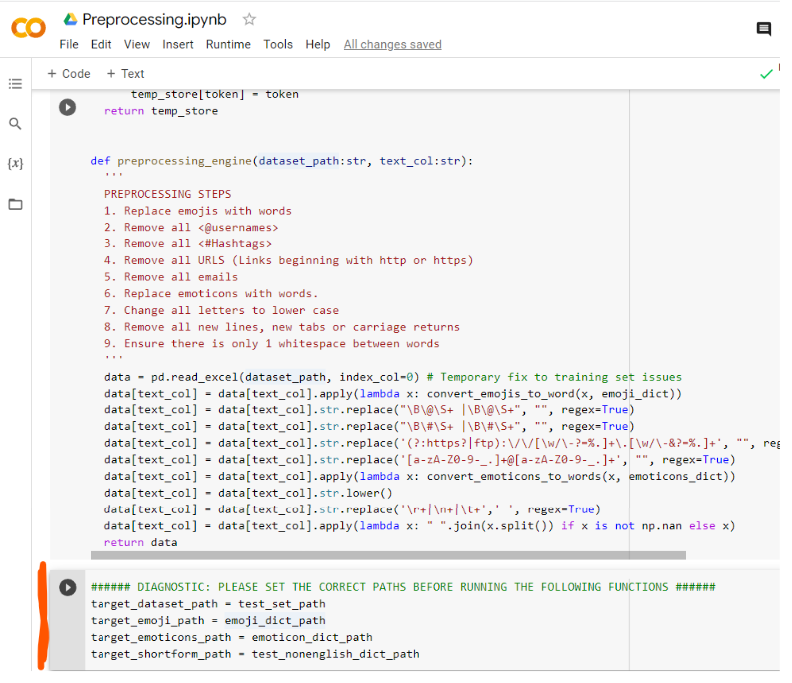
Item 1 originates from the text data collected internally and should be uploaded into the *assets* folder. It requires the text and label columns.

**Getting Started**

Make a copy of entire folder to own drive

In code block #8, change the paths accordingly:

* target \_dataset\_path: The path of the dataset where you want to extract emojis, emoticons and abbreviations
* target\_emoji\_path: The path of the emojis excel file
* target\_emoticons\_path: The path of the emoticons excel file
* target\_shortform\_path: The path of the abbreviations excel file



*Paths to be changed in code block #4*

Change to a GPU runtime in the ribbon under the *Runtime > Change runtime type* button

After successfully connecting to a GPU, click on *Runtime > Run All*

When code executes code block #3, there would be a window pop-up asking for permission to allow Colab to read and write files to your Google Drive, choose your account and allow the necessary permissions

Code block #12 is the slowest step of all the blocks, The rough runtime to execute code block #12 is roughly 4 to 5 hours. This is due to the scraping technology used by the external library to check if a word is an abbreviation.

All code blocks will execute sequentially and the result file will be located in the runtime files after the final code block is executed

After all code blocks have been successfully executed, check the path where the excel files reside and all new emojis, emoticons and abbreviations would have been added to the excel files.

**Note**

We have provided a preliminary corpus excel of emojis, emoticons and abbreviations for convenience.