Manual for PANGKAT: A Dedicated Tokenizer for the Tagalog Language

PANGKAT was designed to be implemented as a Python module, which can be imported into more advanced Tagalog NLP applications as its tokenizer.

```
module > main.py > ...
  1   import pangkat
  2
  3   # Initialize PANGKAT, as it is implemented as a Python Class
  4   PANGKAT = pangkat.PANGKAT()
  5
  6   # To use PANGKAT, call its labelTokens function
  7   # Its parameter is the file containing the data to be tokenized
  8   # The labelTokens function returns the arrays of the resulting tokens and their labels
  9   # Results are stored in independent arrays for both short and longer unit tokenization
 10   tokenList, labelList, longerTokenList, longerLabelList = PANGKAT.labelTokens("input.txt")
```

For instance, the results of PANGKAT are as follows for tokenizing the following sentence:

“MANILA, Philippines — Pinababalik ni Hon. Antonio Kho, Jr. and P60 billion na budget ng Philippine Health Insurance Corporation (PhilHealth) matapos makwestiyon ang malaki-laking budget nito na 'di umano'y hindi napapakinabangan ng publiko.”

```
PS C:\Users\Justin\Desktop\PANGKAT\module> python main.py
PANGKAT is loaded

Short Unit Tokenization Results:

[['MANILA', ',', 'Philippines', '—', 'Pinababalik', 'ni', 'Hon', '.', 'Antonio', 'Kho', ',', 'Jr', '.', 'and', 'P60'
, 'billion', 'na', 'budget', 'ng', 'Philippine', 'Health', 'Insurance', 'Corporation', '(', 'PhilHealth', ')', 'mata
pos', 'makwestiyon', 'ang', 'malaki', '-', 'laking', 'budget', 'nito', 'na', "'", 'di', 'umano', "'", 'y', 'hindi',
'napapakinabangan', 'ng', 'publiko', '.']]
[['B-LOC', 'I', 'I', 'O', 'O', 'O', 'B-PER', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'B-MWE', 'I', 'O', 'O', 'O', 'B-ORG'
, 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'B-MWE', 'I', 'I', 'O', 'O', 'O', 'B-MWE', 'I', 'I', 'I', 'I', 'O', '
O', 'O', 'O', 'O']]

Longer Unit Tokenization Results:

[['MANILA,Philippines', '—', 'Pinababalik', 'ni', 'Hon.Antonio Kho,Jr.', 'and', 'P60 billion', 'na', 'budget', 'ng',
 'Philippine Health Insurance Corporation (PhilHealth)', 'matapos', 'makwestiyon', 'ang', 'malaki-laking', 'budget',
 'nito', 'na', "'di umano'y", 'hindi', 'napapakinabangan', 'ng', 'publiko', '.']]
[['NE-LOC', 'W', 'W', 'W', 'NE-PER', 'W', 'MWE', 'W', 'W', 'W', 'NE-ORG', 'W', 'W', 'W', 'MWE', 'W', 'W', 'W', 'MWE'
, 'W', 'W', 'W', 'W', 'W']]
```

For better visualization, the following images present the results using PANGKAT's graphical user interface for both short and longer unit tokenization, respectively: