# Using Sylbreak in Jupyter Notebook

ဒီ Jupyter Notebook က GitHub မှာ ကျွန်တော်တင်ပေးထားတဲ့ Sylbreak Python ပရိုဂရမ်
https://github.com/ye-kyaw-thu/sylbreak/blob/master/python/sylbreak.py (https://github.com/ye-kyaw-thu/sylbreak/blob/master/python/sylbreak.py) ကို Jupyter Notebook, Python 3 Kernel မှာ function တစ်ခုအနေနဲ့ဆောက်ပြီး သုံးတဲ့ပုံစံကို နမူနာအနေနဲ့ ပြသထားတာ ဖြစ်ပါတယ်။

"sylbreak.py" ကို သုံးစဉ်က input လုပ်တဲ့ စာကြောင်းမှာပါတဲ့ "space" ကို အရင်ဖြတ်ပြီးတော့ (e.g. cat input.txt | sed 's/ //g' | python sylbreak.py ...) သုံးတဲ့ပုံစံနဲ့ သွားခဲ့ပေမဲ့၊ ဒီ sylbreak function မှာတော့ function အတွင်းက ဖြတ်ပေးတဲ့ပုံစံနဲ့ ရေးပြထားပါတယ်။
အဲဒီအတွက် "line = re.sub(ur"\s+","", line)" ဆိုတဲ့ statement တစ်ကြောင်းကို ဖြည့်ရေးထားပါတယ်။

In [1]:
```
# Regular Expression Python Library ကို သုံးလို့ရအောင် import လုပ်တာ
import re

# စာလုံးတွေကို အုပ်စုဖွဲ့တာ (သို့) variable declaration လုပ်တာ
# တကယ်လို့ syllable break လုပ်တဲ့ အခါမှာ မြန်မာစာလုံးချည်းပဲ သပ်သပ် လုပ်ချင်တာဆိုရင် enCh
myConsonant = "က-အ"
enChar = "a-zA-Z0-9"
otherChar = "ဣဤဥဦဧဩဪ၌-၏၀-၉။!-/:-@[-`{-~\s"
ssSymbol = '္'
ngaThat = 'င်'
aThat = 'ိ'

# Regular expression pattern for Myanmar syllable breaking
# *** a consonant not after a subscript symbol AND
# a consonant is not followed by a-That character or a subscript sym
# မြန်မာစာကို syllable segmentation လုပ်ဖို့အတွက်က ဒီ RE pattern တစ်ခုတည်းနဲ့ အဆင်
BreakPattern = re.compile(r"((?<!" + ssSymbol + r")["+ myConsonant +

# sylbreak function ဆောက်တဲ့ အပိုင်း
def sylbreak(line):
        line = re.sub(r"\s+","", line)
        line = BreakPattern.sub(r" " + r"\1", line)
        return line
```

In [2]:
```
# sylbreak function ကိုခေါ်သုံးကြည့်ရအောင်

sylbreak("မြန်မာစာသည် တို့စာ။ တို့စာကို သုတေသန လုပ်ပါ။")
```

Out[2]: ' မြန် မာ စာ သည် တို့ စာ ။ တို့ စာ ကို သု တေ သ န လုပ် ပါ ။'

စိတ်ထဲမှာ ပေါ်လာတာကို ကောက်ရေးပြီးတော့ syllable segmentation လုပ်ခိုင်းလိုက်တာပါ။ :)

နောက်ထပ် ဥပမာအနေနဲ့ Wikipedia Myanmar မှာရေးထားတဲ့ အာခီမီးဒီးစ် ရဲ့ အတ္ထုပ္ပတ္တိအကျဉ်း (https://my.wikipedia.org/wiki/%E1%80%A1%E1%80%AC%E1%80%81%E1%80%AE%E1%80%9
ထဲမှာရေးထားတဲ့
စာကြောင်းတွေကို sylbreak နဲ့ ဖြတ်ကြည့်ရအောင်။

In [3]:
```
sylbreak("""အာခီမီးဒီးစ်ကို ဘီစီ ၂၈၇ ခန့်က ရှေးဟောင်း မဂ္ဂနာဂရေစီယာပြည်လက်အောက်ခံ စစ္စလီပြည်

""")
```

Out[3]:
```
' အာ ခီ မီး ဒီး စ် ကို ဘီ စီ ၂ ၈ ၇ ခ နှ် က ရှေး ဟောင်း မဂ္ဂ နာ ဂ ရေ စီ ယာ ပြည် လက် အောက် ခံ စစ္စ လီ ပြည် ဆိုင် ရာ ကျွှး စ် မြို့ တွင် မွေး ဖွား ခဲ့ သည် ။ ဘိုင် ဇန် တိုင်း ဂ ရိ ခေတ် က သ မိုင်း ပ ည ာ ရှင် ဂျွန် ဇီ ဇီ ၏ မှတ် တမ်း အ ရ အာ ခီ မီး ဒီး စ် သည် အ သက် ၇ ၅ နှစ် အ ထိ နေ ထိုင် သွား ရ ကြောင်း သိ ရ သည် ။ အာ ခီ မီး ဒီး စ် သည် သူ ၏ တီ ထွင် မှု တစ် ခု ဖြစ် သော သဲ နာ ရီ နှ င့် ပတ် သက် ၍ ရေး သား ထား သော S a n d R e c k o n e r s အ မည် ရှိ စာ တမ်း များ တွင် သူ ၏ ဖ ခင် အ မည် ကို နက္ခတ္တ ဗေ ဒ ပ ညာ ရှင် ဖီ း ဒီ း ယပ်စ် ဟု ဖော် ပြ ထား သည် ။ သ မိုင်း ပ ည ာ ရှင် ပ လူး တပ် ရေး သား သော ခေတ် ပြိုင် ပုဂ္ဂိုလ် ထူး ကြီး များ စာ အုပ် တွင် အာ ခီ မီး ဒီး စ် သည် ဆိုင် ရာ ကျွှး စ် ဘု ရင် ဒု တိ ယ မြောက် ဟီ ရိ း နှ င့် ဆွေ မျိုး တော် စပ် ကြောင်း ဖော် ပြ ထား သည် ။ သူ ငယ် ရွယ် စဉ် က အီ ဂျစ် ပြည် အ လက် ဇန္ဒြီး ယား မြို့ တွင် ပ ညာ ဆည်း ပူး ခဲ့ သည် ဟု ယူ ဆ ရ သည် ။ ဘီ စီ ၂ ၁ ၂ တွင် အာ ခီ မီး ဒီး စ် သေ ဆုံး ခဲ့ သည် ။ ရော မ စစ် ဗိုလ် ချုပ် မား ကပ်စ် က လ ော ဒီး ယပ်စ် မာ ဆဲ လပ်စ် က နှစ် နှစ် ကြာ ဝိုင်း ရံ ပိတ် ဆို့ ပြီး နောက် ဆိုင် ရာ ကျွှး စ် မြို့ ကို သိမ်း ပို က် လိုက် သည် ။ ထို အ ချိန် တွင် အာ ခီ မီး ဒီး သည် ကျော် မက် ထ ရိ ပုစ္ဆာ တစ် ပုဒ် ကို စဉ်း စား အ ဖ ြေ ရှာ နေ ခိုက် ဖြစ် သည် ။ ရော မ စစ် သား က သူ့ အား ဖမ်း ဆီး လိုက် ပြီး ဗိုလ် ချုပ် မာ ဆဲ လပ်စ် နှ င့် တွေ့ ဆုံ ရန် ပြော ဆို ရာ သူ က သူ ၏ ပုစ္ဆာ စဉ်း စား နေ ဆဲ ဖြစ် ၍ မ တွေ့ လို ကြောင်း ငြင်း ဆို သည် တွင် ရော မ စစ် သား က ဒေါ သ ထွက် ကာ မား ဖြ ့ င့် ထိုး သတ် လိုက် သည် ဟု ပ လူး တပ် က ရေး သား ခဲ့ သည် ။ ဗိုလ် ချုပ် မာ ဆဲ လပ်စ် သည် အာ ခီ မီး ဒီး စ် သေ ဆုံး သွား သ ည့် အ တွက် များ စ ွာ နှ မြော တ သ ဖြစ် ရ သည် ။ အာ ခီ မီး ဒီး စ် အား ပ ညာ ရှင် တစ် ယောက် အ ဖြစ် သိ ရှိ ထား သ ော ကြော ့ င် မ သတ် ရန် ကြိ တင် အ မိ နှ် ပေး ထား ခဲ့ သည် ။“ ငါ့ စက် ဝိုင်း တွေ ပေါ် တက် မ နင်း ပါ နဲ့ ” ဟူ သော စ ကား ကို အာ ခီ မီး ဒီး စ် နောက် ဆုံး ပြော ဆို ခဲ့ သည် ဟု အ ချို့ က ယူ ဆ ကြ သ ော် လည်း သ မိုင်း ပ ညာ ရှင် ပ လူး တပ် ရေး သော စာ အုပ် တွင် မူ မ ပါ ရှိ ပေ ။ အာ ခီ မီး ဒီး စ် ၏ ဂူ ဗိမ္မာန် တွင် ထု လုံး ရှည် မှန် တစ် ခု အ တွင်း စက် လုံး တစ် ခု ကို ထ ည့် သွင်း ထား သ ည့် ရုပ် တု တစ် ခု ကို စိုက် ထူ ထား သည် ။ အာ ခီ မီး ဒီး စ် သေ ဆုံး ပြီး နှစ် ပေါင်း ၁ ၃ ၇ နှစ် အ ကြာ ဘီ စီ ၇ ၅ တွင် ရော မ ခေတ် နိုင် ငံ ရေး သု ခ မိန် ဆီ ဇာ ရိ က အာ ခီ မီး ဒီး စ် အ ကြောင်း ကြား သိ ရ ၍ သူ ၏ အုတ် ဂူ အား ရှာ ဖွေ ခဲ့ သည် ။ ခြေ နှ ယ် ပိတ် ပေါင်း များ ဖုံး အုပ် နေ သော အာ ခီ မီး ဒီး စ် ၏ အုတ် ဂူ ကို ဆိုင် ရာ ကျွှး စ် မြို့ အ နီး တွင် ရှာ ဖွေ တွေ့ ရှိ ခဲ့ ပြီး သ နှ့် ရှင်း ရေး ပြု လုပ် ကာ အုတ် ဂူ ပေါ် မှ စာ သား များ ကို ဖတ် ရှု သွား သည် ။ ဆိုင် ရာ ကျွှး စ် စစ် ပွဲ အ ပြီး နှစ် ပေါင်း ၇ ၀ အ ကြာ တွင် ပို လီ း ဘီ း ယပ်စ် ရေး သား သော ဆိုင် ရာ ကျွှး စ် စစ် ပွဲ အ ကြောင်း စာ အုပ် တွင် အာ ခီ မီး ဒီး စ် နှ င့် ပ တ် သက် သော အ ကြောင်း များ ပါ ရှိ ၍ သ မိုင်း ပ ညာ ရှင် ပ လူး တပ် က ထပ် မံ ရေး သား နှိုင် ခဲ့ ခြ င်း ဖြစ် ပါ သည် ။ ဆိုင် ရာ ကျွှး စ် မြို့ က ကွယ် ရေး အ တွက် စစ် ပွဲ ဝင် စက် ကို ရိ ယာ လက် နက် ဆန်း များ ကို လည်း အာ ခီ မီး ဒီး စ် က တီ ထွင် ပေး ခဲ့ ကြောင်း အ ဆို ပါ စာ အုပ် တွင် ဖော် ပြ ပါ ရှိ ပါ သည် ။ '
```

# Typing order

မြန်မာစာနဲ့ ပတ်သက်တဲ့ NLP (Natural Language Processing) အလုပ် တစ်ခုခု လုပ်ဖို့အတွက် syllable segmentation လုပ်ကြမယ်ဆိုရင် တကယ်တမ်းက မလုပ်ခင်မှာ၊ မြန်မာစာ စာကြောင်းတွေရဲ့ typing order အပါအဝင် တခြား ဖြစ်တတ်တဲ့ အမှားတွေကိုလည်း cleaning လုပ်ရပါတယ်။ အဲဒီလိုမလုပ်ရင် sylbreak က ကျွန်တော် အကြမ်းမျဉ်း သတ်မှတ်ထားတဲ့ မြန်မာစာ syllable unit တွေအဖြစ် မှန်မှန်ကန်ကန် ဖြတ်ပေးနိုင်မှာ မဟုတ်ပါဘူး။ မြန်မာစာ စာကြောင်းတွေထဲမှာ ရှိတတ်တဲ့အမှား တွေက တကယ့်ကို အများကြီးပါ။ တချို့ အမှားတွေက မျက်လုံးနဲ့ကြည့်ယုံနဲ့ မခွဲခြားနိုင်တာမျိုးတွေလည်း ရှိပါတယ်။ ဒီနေရာမှာတော့ အမှားအမျိုးအစားတွေထဲက တစ်မျိုးဖြစ်တဲ့ typing order အမှား တစ်မျိုး၊ နှစ်မျိုးကို ဥပမာအနေနဲ့ရှင်းပြရင်း၊ အဲဒီလိုအခြေအနေမျိုးမှာ ဖြစ်တတ်တဲ့ sylbreak က ထွက်လာမယ့် အမှား output တွေကိုလည်း လေ့လာကြည့်ကြရအောင်။

အောက်မှာ သုံးပြထားတဲ့ "ခန့်" က "ခ န ့ ်" (ခခွေး နငယ် အောက်မြစ် အသတ်) ဆိုတဲ့ အစီအစဉ် အမှားနဲ့ ရိုက်ထားတာဖြစ်ပါတယ်။ အဲဒါကြောင့် sylbreak က ထွက်လာတဲ့အခါမှာ "ခခွေး" နဲ့ "နငယ် အသတ် အောက်မြစ်" က ကွဲနေတာဖြစ်ပါတယ်။

In [4]:
```
1 sylbreak("ဘီစီ ၂၈၇ ခန့်")
```

Out[4]: ' ဘီ စီ ၂ ၈ ၇ ခ နှ့်'

တကယ်တန်း မှန်ကန်တဲ့ "ခန့်" ရဲ့ typing order က "ခ န ် ့" (ခခွေး နငယ် အသတ် အောက်မြစ်) ပါ။
အမြင်အားဖြင့်ကတော့ မခွဲနိုင်ပေမဲ့၊ မှန်ကန်တဲ့ typing order နဲ့ ရိုက်ထားရင်တော့ "ခန့်" ဆိုပြီး syllable တစ်ခုအနေနဲ့ ရိုက်ထုတ်ပြပေးပါလိမ့်မယ်။

In [5]:
```
1 sylbreak("ဘီစီ ၂၈၇ ခန့်")
```

Out[5]: ' ဘီ စီ ၂ ၈ ၇ ခန့်'

နောက်ထပ် typing order အမှားတစ်ခုကို ကြည့်ကြရအောင်။

In [6]:
```
1 sylbreak("ထည့်သွင်းထားသည့်ရုပ်တု")
```

Out[6]: ' ထ ည့် သွင်း ထား သ ည့် ရုပ် တု'

"ညကြီး အောက်မြစ် အသတ်" ဆိုတဲ့ မှားနေတဲ့ အစီအစဉ်ကို "ညကြီး အသတ် အောက်မြစ်" ဆိုပြီး ပြောင်းရိုက်ပြီးတော့ sylbreak လုပ်ကြည့်ရင်တော့ အောက်ပါအတိုင်း "ထ" နဲ့ "ည့်", "သ" နဲ့ "ည့်" တွေက ကွဲမနေတော့ပဲ မှန်မှန်ကန်ကန်ဖြတ်ပေးပါလိမ့်မယ်။

In [7]:
```
1 sylbreak("ထည့်သွင်းထားသည့်ရုပ်တု")
```

Out[7]: ' ထည့် သွင်း ထား သည့် ရုပ် တု'

တချို့အမှားတွေကတော့ ဂရုစိုက်ရင် မျက်စိနဲ့ မြင်နိုင်ပါတယ်။
ဥပမာ "ဥ" (အကွရာ ဥ) နဲ့ "ဉ" (ညကလေး) ကိုမှားရိုက်တဲ့ကိစ္စပါ။

သို့သော် ကျွန်တော်မြန်မာစာကြောင်းတွေအများကြီးကို ကိုင်တွယ်အလုပ်လုပ်တဲ့အခါတိုင်းမှာ ဒီလိုအမှားက အမြဲတမ်းကို ပါတတ်ပါတယ်။
ဖောင့် (font) မှာလည်း မှန်မှန်ကန်ကန်ခွဲထားမယ်ဆိုရင်၊ အမှန်က ညကလေးဆိုရင် အမြီးက ရှည်ပါတယ်။ စာရိုက်သူအများစုက သတိမပြုမိတဲ့ အကြောင်းအရင်း တစ်ခုကလည်း တချို့ text editor တွေမှာ "အကွရာ ဥ" နှင့် ညကလေး "ဉ" ကို ကွဲပြားအောင် မပြသပေးနိုင်လို့ပါ။

In [8]:
```
1 sylbreak("ကာရီသည်ဒီနှစ်၏ပါရမီရှင်တစ်ဦးနှင့်ထိုက်တန်သောအမျိုးသမီးအဆိုရှင်ဖြစ်သည်။")
```

Out[8]: ' ကာ ရီ သည် ဒီ နှစ် ၏ ပါ ရ မီ ရှင် တစ် ဦး နှ င့် ထိုက် တန် သော အ မျိုး သ မီး အ ဆို ရှင် ဖြစ် သ ည် ။'

ဝီကီပီးဒီးယားက မှားနေတဲ့ "ညကလေး" ကို "အကွရာ ဥ" နဲ့ပြန်ပြင်ရိုက်ထားတဲ့ စာကြောင်းနဲ့ နောက်တစ်ခေါက် syllable ဖြတ်ထားတာက အောက်ပါအတိုင်းဖြစ်ပါတယ်။ "ညကလေး" နဲ့ "အကွရာ ဥ" အမှားကိစ္စမှာတော့ syllable segmentation ဖြတ်တဲ့အပိုင်းမှာတော့ ထူးထူးခြားခြား အပြောင်းအလဲ မရှိပါဘူး။

In [9]:
```
1 sylbreak("ကာရီသည်ဒီနှစ်၏ပါရမီရှင်တစ်ဦးနှင့်ထိုက်တန်သောအမျိုးသမီးအဆိုရှင်ဖြစ်သည်။")
```

Out[9]: ' ကာ ရီ သည် ဒီ နှစ် ၏ ပါ ရ မီ ရှင် တစ် ဦး နှ င့် ထိုက် တန် သော အ မျိုး သ မီး အ ဆို ရှင် ဖြစ် သ ည် ။'

# Note

- sylbreak မှာ သုံးထားတဲ့ မြန်မာစာ syllable unit (အဖြတ်အတောက်တွေ) က ကျွန်တော်လုပ်ခဲ့တဲ့ NLP (Natural Language Processing) သုတေသနအလုပ်တွေဖြစ်တဲ့ Machine Translation, Automatic Speech Recognition, Text to Speech, POS tagging စတဲ့ အလုပ်တွေအတွက် လုပ်ရကိုင်ရ အဆင်အပြေဆုံး ပုံစံအတိုင်း တကယ့်ကို simple unit အနေနဲ့ဖြတ်ထားတာ ဖြစ်ပါတယ်။ အဲဒါကြောင့် "ဘီစီ ၂၈၇" ကို "ဘီ စီ ၂ ၈ ၇"၊ "နက္ခတ္တ ဗေဒ" ကို "နက္ခတ္တ ဗေ ဒ"၊ နောက်ပြီးတော့ မြန်မာစာတွေနဲ့ အတူတူရောပါနေတဲ့ "Sand Reckoners" ကို "S a n d R e c k o n e r s" ဆိုပြီး ဖြတ်ထားပါတယ်။ ဆိုလိုတာက ပါဠိဆင့်တွေကို ဖြေတဲ့ အလုပ် (ဥပမာ နက္ခတ္တ ကို နက် ခတ် တ)၊ ဂဏန်းစာလုံးတွေကို တွဲတဲ့အလုပ် (ဥပမာ ၂၈၃)၊ ရောပါနေတဲ့ အင်္ဂလိပ်စာလုံးတွေကို ဖယ်ပစ်တာ၊ နဂိုအတိုင်းပဲ တွဲထားတာ (ဥပမာ Sand Reckoners) မျိုးတွေကို တမင်တကာ လုပ်မထားတာပါ။ အကြောင်းအရင်းကတော့ အမျိုးမျိုးရှိပါတယ်။ ဥပမာ ပါဠိဆင့်တွေကို ဖြေပစ်လိုက်ရင် လိုအပ်တဲ့အခါမှာ နဂိုပုံစံအတိုင်းပြန်ရအောင် ပြန်ဆင့်ရပါတယ်။ အဲဒီအလုပ်က လွယ်မလိုလိုနဲ့ လက်တွေ့မှာတော့၊ ပါဠိဆင့် ပြန်ဆင့်ပေးရတဲ့အလုပ်အတွက် processing time နဲ့ အဲဒီကနေထွက်လာမဲ့ error တွေကို ရှာဖွေရတဲ့အလုပ်၊ ပြန်ပြင်ပေးရတဲ့အလုပ်တွေကို ရှောင်ချင်လို့ ဖြစ်ပါတယ်။ sylbreak (https://github.com/ye-kyaw-thu/sylbreak) က ကျွန်တော်တို့ မြန်မာစာကို unicode နဲ့သာ မှန်မှန်ကန်ကန်ရေးထားရင်၊ "Regular Expression တစ်ကြောင်းထဲနဲ့ လွယ်လွယ်ကူကူ ဖြတ်လို့ရကြောင်း" နောက်ပြီးတော့ အဲဒါက "တကယ်လည်း မြန်မာစာ NLP သုတေသနအလုပ်တွေအတွက် အသုံးဝင်ကြောင်း"၊ ဒါ့အပြင် "အခြေခံကြတဲ့ မြန်မာဝဏ္ဏ (syllable unit) အနေနဲ့လည်း ရပါလိမ့်မယ်" ဆိုတဲ့ message ကိုပေးထားတာပဲ ဖြစ်ပါတယ်။ ကျွန်တော်ရဲ့ syllable unit တွေက မြန်မာစာအနေနဲ့ကြည့်ရင် ပြည့်စုံမှန်ကန်တယ်လို့ မဆိုလိုပါဘူး။ ကိုယ်လုပ်မဲ့ အလုပ်၊ develop လုပ်နေတဲ့ application ပေါ်ကို မူတည်ပြီးတော့ လက်ရှိ ကျွန်တော်ပြင်ဆင်ပေးထားတဲ့ Regular Expression ကို ကြိုက်သလို ဖြည့်စွက်တာ၊ ပြင်သုံးတာကိုလုပ်နိုင်ပါတယ်။
- Python 3.4 ကနေစပြီးတော့ "ur" (Unicode + Raw text) ဆိုပြီးတွဲရေးတာကို support မလုပ်ပါဘူး။ သို့သော် "u" တစ်လုံးတည်း "r" တစ်လုံးတည်း သုံးတာကိုတော့ ခွင့်ပြုပါတယ်။

https://stackoverflow.com/questions/26063899/python-version-3-4-does-not-support-a-ur-prefix (https://stackoverflow.com/questions/26063899/python-version-3-4-does-not-support-a-ur-prefix)

- Jupyter Notebook နဲ့ ပတ်သက်တဲ့ installation လုပ်ပုံလုပ်နည်း၊ အသုံးပြုပုံနဲ့ ပတ်သက်ပြီး မြန်မာလိုလေ့လာချင်တဲ့ သူများအတွက် ကျွန်တော့်ရဲ့ Tutorial (https://github.com/ye-kyaw-thu/Tutorials) မှာ လေ့လာနိုင်ပါတယ်။