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 တစ်ခုအနေနဲ့ ဆောက် ပြီး 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 ရေးပြထားပါ တယ်။

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 enCha
         r က မလိုပါဘူး
         myConsonant = "ന-ജ"
         enChar = "a-zA-Z0-9"
         otherChar = "ဣဤဥဦီဪဿ၌၍၏၀-၉။။!-/:-@[-`{-~\s"
ssSymbol = '့'
         ngaThat = '\delta'
         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 symb
         οl
         # မြန်မာစာကို syllable segmentation လုပ်ဖို့အတွက်က ဒီ RE pattern တစ်ခုတည်းနဲ့ အဆင်
         ပြေတယ်။
         BreakPattern = re.compile(r"((?<!" + ssSymbol + r")["+ myConsonant +</pre>
         r"](?![" + aThat + ssSymbol + r"])" + r"|[" + enChar + otherChar + r"]
         "])", re.UNICODE)
         # 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။")
```

 $\mathsf{Out}[2]$: 'မြန် မာ စာ သည် တို့ စာ ။ တို့ စာ ကို သု တေ သ န လုပ် ပါ ။'

စိတ်ထဲမှာ ပေါ် လာတာကို ကောက်ရေးပြီးတော့ syllable segmentation လုပ်ခိုင်းလိုက်တာပါ။ :)

နောက်ထပ် ဥပမာအနေနဲ့ <u>Wikipedia Myanmar မှာရေးထားတဲ့ အာခီမီးဒီးစ် ရဲ့ အတ္ထုပ္ပတ္တိအကျဉ်း</u> (<u>https://my.wikipedia.org/wiki/%E1%80%A1%E1%80%AC%E1%80%81%E1%80%AE%E1%80%99%E1%80%AE</u> ထဲမှာရေးထားတဲ့

စာကြောင်းတွေကို sylbreak နဲ့ ဖြတ်ကြည့်ရအောင်။

In [3]:

sylbreak (" " " အာခ်ီမီးဒီးစ်ကို ဘီစီ ၂၈၇ ခန့်က ရေးဟောင်း မဂ္ဂနာဂရေစီယာပြည်လက်အောက်ခံ စစ္စလီပြ ည် ဆိုင်ရာကျူးစ် မြို့ တွင် မွေးဖွားခဲ့သည်။ ဘိုင်ဇန်တိုင်းဂရိခေတ် က သမိုင်းပညာရှင် ဂျွန်ဇီဇီ ၏ မှတ်တမ်းအရ အာဓီမီးဒီးစ်သည် အသက် ၇၅ နှစ်အထိ နေထိုင်သွားရကြောင်း သိရသည်။ အာဓီမီးဒီးစ်သည် သူ၏ တီထွင်မှု တ စ်ခုဖြစ်သော သဲနာရီ နှင့် ပတ်သက်၍ ရေးသားထားသော Sand Reckoners အမည်ရှိ စာတမ်းများတွင် သူ ၏ ဖခင်အမည်ကို နက္ခတ္တဗေဒပညာရှင် ဖီးဒီးယပ်စ် ဟု ဖော်ပြထားသည်။ သမိုင်းပညာရှင် ပလူးတပ် ရေးသား သော ခေတ်ပြိုင်ပုဂ္ဂိုလ်ထူးကြီးများ စာအုပ်တွင် အာခီမီးဒီးစ်သည် ဆိုင်ရာကျူးစ်ဘုရင် ဒုတိယမြောက်ဟီရိုးနှင့် ဆွေ မျိုး တော်စပ်ကြောင်း ဖော်ပြထားသည်။ သူငယ်ရွယ်စဉ်က အီဂျစ်ပြည် အလက်မန္ဒြီးယားမြို့ တွင် ပညာဆည်းပူး ့ခဲ့သည်ဟု ယူဆရသည်။ ဘီစီ ၂၁၂ တွင် အာခီမီးဒီးစ် သေဆုံးခဲ့သည်။ ရောမစစ်ဗိုလ်ချုပ် မားကပ်စ် ကလောဒီးယ ပ်စ် မာဆဲလပ်စ် က နှစ်နှစ်ကြာဝိုင်းရံ ပိတ်ဆို့ပြီးနောက် ဆိုင်ရာကျူးစ် မြို့ကို သိမ်းပိုက်လိုက်သည်။ ထိုအချိန်တွင် အာခီမီးဒီးသည် ဂျော်မက်ထရီ ပုစ္ဆာတစ်ပုဒ်ကို စဉ်းစား အဖြေရှာနေခိုက် ဖြစ်သည်။ ရောမစစ်သားက သူ့အား ဖမ်း ဆီးလိုက်ပြီး ဗိုလ်ချုပ် မာဆဲလပ်စ် နှင့် တွေ့ဆုံရန် ပြောဆိုရာ သူက သူ၏ပုစ္ဆာစဉ်းစားနေဆဲဖြစ်၍ မတွေ့လိုကြောင်း ငြင်းဆိုသည်တွင် ရောမစစ်သားကူ ဒေါ့သထွက်ကာ ဓားဖြင့် ထိုးသတ်လိုက်သည်ဟု ပလူးတပ် က ရေးသားခဲ့သည့် ။ ဗိုလ်ချုပ် မာဆဲလပ်စ်သည် အာခီမီးဒီးစ် သေဆုံးသွားသည့် အတွက် များစွာ နှမြောတသဖြစ်ရသည်။ အာခီမီးဒီးစ် အား ပညာရှင် တစ်ယောက်အဖြစ် သိရှိထားသောကြောင့် မသတ်ရန် ကြိုတင် အမိန့်ပေးထားခဲ့သည်။ "ငါ့စက်ဝိုင်း တွေပေါ် တက်မနင်းပါနဲ့ "ဟူသော စကားကို အာခီမီးဒီးစ် နောက်ဆုံး ပြောဆိုခဲ့သည်ဟု အချို့က ယူဆကြသော်လ ည်း သမိုင်းပညာရှင် ပလူးတပ် ရေးသော စာအုပ်တွင်မူ မပါရှိပေ။ အာခ်ီမီးဒီးစ်၏ ဂူဗ်မ္မာန်တွင် ထုလုံးရှည်မှန်တစ် ခုအတွင်း စက်လုံးတစ်ခုကို ထည့်သွင်းထားသည့် ရုပ်တုတစ်ခုကို စိုက်ထူထားသည်။ အာခီမီးဒီးစ် သေဆုံးပြီး နှစ်ပေါ င်း ၁၃၇နှစ်အကြာ ဘီစီ ၇၅တွင် ရောမခေတ် နိုင်ငံရေးသုခမိန် ဆီဇာရိက အာခီမီးဒီးစ် အကြောင်းကြားသိရ၍ သူ ၏ အုတ်ဂူအား ရှာဖွေခဲ့သည်။ ခြုံနွယ်ပိတ်ပေါင်းများ ဖုံးအုပ်နေသော အာခီမီးဒီးစ်၏ အုတ်ဂူကို ဆိုင်ရာကျူးစ်မြို့အ နီးတွင် ရှာဖွေ တွေ့ရှိခဲ့ပြီး သန့်ရှင်းရေးပြုလုပ်ကာ အုတ်ဂူပေါ်မှ စာသားများကို ဖတ်ရှုသွားသည်။ ဆိုင်ရာကျူးစိစ်စ်ပွဲ ့အပြီး နှစ်ပေါင်း ၇၀ အကြာတွင် ပိုလီးဘီးယပ်စ် ရေးသားသော ဆိုင်ရာကျူးစ်စစ်ပွဲ အကြောင်း စာအုပ်တွင် အာခီ မီးဒီးစိနှင့် ပတ်သက်သော အကြောင်းများ ပါရှိ၍ သမိုင်းပညာရှင် ပလူးတပ် က ထပ်မံ ရေးသားနိုင်ခဲ့ခြင်း ဖြစ်ပါသ ည်။ ဆိုင်ရာကျူးစ်မြို့ ကာကွယ်ရေးအတွက် စစ်ပွဲဝင် စက်ကိရိယာ လက်နက်ဆန်းများကိုလည်း အာခီမီးဒီးစ်က တီ ထွင်ပေးခဲ့ကြောင်း အဆိုပါ စာအုပ်တွင် ဖော်ပြပါရှိပါသည်။

""")

Out[3]: ' အာ ခီ မီး ဒီးစ် ကို ဘီ စီ ၂ ၈ ၇ ခ န့် က ရှေး ဟောင်း မဂ္ဂ နာ ဂ ရေ စီ ယာ ပြည် လက် အောက် သွား ရ ကြောင်း သိ ရ သည် ။ အာ ခီ မီး ဒီးစီ သည် သူ ၏ တီ ထွင် မှု တစ် ခု ဖြစ် သော သဲ နာ ရီ နှ ငို ပတိ သက် ၍ ရေး သား ထား သော S a n d R e c k o n e r s အ မည် ရှိ စာ တစ်း မ ျား တွင် သူ ၏ ဖ ခင် အ မည် ကို နက္စတ္တ ဗေ ဒ ပ ညာ ရှင် ဖီး ဒီး ယဝ်စ် ဟု ဖော် ပြ ထား သည် ။ သ မိုင်း ပ ညာ ရှင် ပ လူး တပ် ရေး သား ထာ ခေတ် ပြိုင် ပုဂ္ဂိုလ် ထူး ကြီး များ စာ အုပ် တွင် အာ ခီ မီး ဒီးစ် သည် ဆိုင် ရာ ကျူးစ် ဘု ရင် ဒု တိ ယ မြောက် ဟီ ရိုး နှ င့် ဆွေ မျိုး တော် စပဲ ကြောင်း ဖ ော် ပြ ထား သည် ။ သူ ငယ် ရွယ် စဉ် က အီ ဂျစ် ပြည် အ လက် မြွေး ယား မြို့ တွင် ပ ညာ ဆည်း ပူး ခဲ့ သည် ။ သူ ငယ် ရွယ် စဉ် က အီ ဂျစ် ပြည် အ လက် မြွေး ယား မြို့ တွင် ပ ညာ ဆည်း ပူး ခဲ့ သည် ဟု ယူ ဆ ရ သည် ။ ဘီ စီ ၂ ၁ ၂ တွင် အာ ခီ မီး ဒီးစ် သေ ဆုံး ခဲ့ သည် ။ ရော မ စစ် ဗိုလ် ချုပ် မား ကပ်စ် က လော ဒီး ယပ်စ် မာ ဆဲ လပ်စ် က နှစ် နှစ် ကြာ ပိုင်း ရဲ ပတ် ဆို ပြီး နောက် ဆို င် ရာ ကျူးစ် မြို့ ကို သိမ်း ပိုက် လိုက် သည် ။ ထို အ ချိန် တွင် အာ ခီ မီး ဒီး သည် ရော် မက် ထရီ ပု စွာ တစ် ပုဒ် ကို စဉ်း စား အ ဖြေ ရှာ နေ ခိုက် ဖြစ် သည် ။ ရော မ စစ် သား က သူ အား မမ်း ဆီး လိုက် ပြီး ဗိုလ် ချုပ် မာ ဆဲ လပ်စ် နှ င့် တွေ့ ဆုံ ရန် ပြော ဆို ရာ သည် ။ ရော မ စစ် သည် အာ ခီ မီး ဒီးစ် သည် အာ ခီ မီး ဒီးစ် သည် အာ ခီ မီး သည် တွင် ရော မ စစ် သား က ခေါ သ ထွက် ကာ ဓား ဖြင့် ထိုး သတ် လိုက် သည် ပု ပ လူး တပ် က ရေး သား ခဲ့ သည် ။ ရော မ စစ် သား က အ ခါ မီး ဒီးစ် အား ပ ညာ ရှင် တစ် သောက် သည့် အာ ခီ မီး ဒီးစ် အား ပ ညာ ရှင် တက် လိုက် သည် ဟု သူ အာ အကို များ စွာ နှ မြော တ သ ပြစ် ချည် ။ အာ ခီ မီး ဒီးစ် အား ပ ညာ ရှင် တက် လိုင်း တွင် ပေ၊ တက် မနင်း ပါ နဲ့ " ဟု သော စ ကား ကို အာ ခီ မီး ဒီးစ် နေး သာ စာ အုပ် တွင် ရာ မ တစ် ကောက် သည် ။ အာ ခီ မီး ဒီးစ် ဆေ သည် ။ ရှင် သည် ။ အစ် ခစ် မီး ဒီးစ် ဆေ ဆည် ဖ ကြာသင်း တား သ ညှိ ရာ ကွက် များ တွင် ရာ မ တစ် လည် အာ ခီ မီး ဒီးစ် အား ပ ညာ ရှင် စ ကို လို ကို တု တစ် မေးသာ စာ အုပ် တိုင်း စာ သည် ။ အာ ခီ မီး ဒီးစ် ဆေ တာ သ မိစ် သည် ။ အာ ခီ မီး ဒီးစ် ဆေ တာ သ မိစ် သည် အာ ခီ မီး ဒီးစ် ဆေ ကွောင်း မောင် လုပ် သည် ။ ခု ခို စည် လှစ် ရက် လှစ် ရက် လှစ် ရက် လှစ် ရာ တွင် ရာ မ မေတိ နဲ့ ပြီး နေ ပေင်း ၁ ၃ ၇ နှစ် အကြာသင်း ကြား သိ ရ ၅ ၅ တွင် ရော မ စေတိ နှစ် ငံ ရေး သာ ခ မိန် ဆီ စာ ရိုင်း ရာ ကျေးစ် စဲ နဲ့ အင်း နေ သော အခု မီး ဒီးစ် ကေ ကြောင်း မော

Typing order

မြန်မာစာနဲ့ ပတ်သက်တဲ့ NLP (Natural Language Processing) အလုပ် တစ်ခုခု လုပ်ဖို့အတွက် syllable segmentation လုပ်ကြမယ် ဆိုရင် တကယ်တမ်းက မလုပ်ခင်မှာ၊ မြန်မာစာ စာကြောင်းတွေရဲ့ typing order အပါအဝင် တခြား ဖြစ်တတ်တဲ့ အမှားတွေကိုလည်း cleaning လုပ်ရပါတယ်။ အဲဒီလိုမလုပ်ရင် sylbreak က ကျွန်တော် အကြမ်းမျဉ်းသတ်မှတ်ထားတဲ့ မြန်မာစာ syllable unit တွေအဖြစ် မှန်မှန်ကန်ကန် ဖြတ်ပေးနိုင်မှာ မဟုတ်ပါဘူး။ မြန်မာစာ စာကြောင်းတွေထဲမှာ ရှိတတ်တဲ့အမှား တွေက တကယ့်ကို အများကြီးပါ။ တချို့ အမှားတွေက မျက်လုံးနဲ့ ကြည့်ယုံနဲ့ မခွဲခြားနိုင်တာမျိုးတွေလည်း ရှိပါတယ်။ ဒီနေ ရာမှာတော့ အမှားအမျိုးအစားတွေထဲက တစ်မျိုးဖြစ်တဲ့ typing order အမှား တစ်မျိုး၊ နှစ်မျိုးကို ဥပမာအနေနဲ့ရှင်းပြရင်း၊ အဲဒီလိုအခြေအနေမျိုးမှာ ဖြစ်တတ်တဲ့ sylbreak က ထွက်လာမယ့် အမှား output တွေကိုလည်း လေ့လာကြာ ည်ကြရအောင်။

အောက်မှာ သုံးပြထားတဲ့ "ခန့်" က "ခ န ့ ်" (ခနွေး နငယ် အောက်မြစ် အသတ်) ဆိုတဲ့ အစီအစဉ် အမှားနဲ့ ရိုက်ထားတာဖြစ်ပါတယ်။ အဲဒါကြော င့် Sylbreak က ထွက်လာတဲ့အခါမှာ "ခနွေး" နဲ့ "နငယ် အသတ် အောက်မြစ်" က ကွဲနေတာဖြစ်ပါတယ်။

In [4]: sylbreak("නීစီ ၂၈၇ ခန့်")

Out[4]: 'ဘီစီ၂၈၇ခန့်'

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

In [5]: sylbreak("ဘီစီ ၂၈၇ ခန့်")

Out[5]: 'ဘီစီ ၂၈၇ခန့်'

နောက်ထပ် typing order အမှားတစ်ခုကို ကြည့်ကြရအောင်။

In [6]: sylbreak("ထည့်သွင်းထားသည့်ရုပ်တု")

Out[6]: 'ထည့်သွင်းထားသည့် ရုပ်တု'

"ညကြီး အောက်မြစ် အသတ်" ဆိုတဲ့ မှားနေတဲ့ အစီအစဉ်ကို "ညကြီး အသတ် အောက်မြစ်" ဆိုပြီး ပြောင်းရိက်ပြီးတော့ sylbreak လုပ်ကြည့်ရင် တော့ အောက်ပါအတိုင်း "ထ" နဲ့ "ည့်", "သ" နဲ့ "ည့်" တွေက ကွဲမနေတော့ပဲ မှန်မှန်ကန်ကန်ဖြတ်ပေးပါလိမ့်မယ်။

In [7]: sylbreak ("ထည့်သွင်းထားသည့်ရုပ်တု")

Out[7]: 'ထည့် သွင်း ထား သည့် ရုပ် တု'

တချို့အမှားတွေကတော့ ဂရုစိုက်ရင် မျက်စိနဲ့ မြင်နိုင်ပါတယ်။ ဥပမာ "ဉ" (အက္ခရာ ဉ) 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 "အက္ခရာ ဉ" နှင့် ညကလေး "ဉ" ကို ကွဲပြားအောင် မပြသပေးနိုင်လို့ပါ။

In [8]: 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 သည် ။'

ဝီကီပီးဒီးယားက မှားနေတဲ့ "ညကလေး" ကို "အက္ခရာ ဉ" နဲ့ပြန်ပြင်ရိုက်ထားတဲ့ စာကြောင်းနဲ့ နောက်တစ်ခေါက် syllable ဖြတ်ထားတာက အောက်ပါအတိုင်းဖြစ်ပါတယ်။ "ညကလေး" နဲ့ "အက္ခရာ ဉ" အမှားကိစ္စမှာတော့ syllable segmentation ဖြတ်တဲ့အပိုင်းမှာတော့ ထူးထူးခြား ခြား အပြောင်းအလဲ မရှိပါဘူး။

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

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

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 အနေနဲ့ ဖြတ်ထားတာ 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။ ဆိုလိုတာ က ပါဠိဆင့်တွေကို ဖြေတဲ့ အလုပ် (ဥပမာ နက္ခတ္ထ ကို နက် ခတ် တ)၊ ဂဏန်းစာလုံးတွေကို တွဲတဲ့အလုပ် (ဥပမာ ၂၈၃)၊ ရောပါနေ တဲ့ အင်္ဂလိပ်စာလုံးတွေကို ဖယ်ပစ်တာ၊ နဂိုအတိုင်းပဲ တွဲထားတာ (ဥပမာ Sand Reckoners) မျိုးတွေကို တမင်တကာ လုပ်မထား တာပါ။ အကြောင်းအရင်းကတော့ အမျိုးမျိုးရှိပါတယ်။ ဥပမာ ပါဠ်ဆင့်တွေကို ဖြေပစ်လိုက်ရင် လိုအပ်တဲ့အခါမှာ နဂိုပုံစံအတိုင်းပြန် ရအောင် ပြန်ဆင့်ရပါတယ်။ အဲဒီအလုပ်က လွယ်မလိုလိုနဲ့ လက်တွေ့မှာတော့၊ ပါဌ်ဆင့် ပြန်ဆင့်ပေးရတဲ့အလုပ်အတွက် processing time နဲ့ အဲဒီကနေထွက်လာမဲ့ error တွေကို ရှာဖွေရတဲ့အလုပ်၊ ပြန်ပြင်ပေးရတဲ့အလုပ်တွေကို ရောင်ချင်လို့ ဖြစ်ပါတယ်။ <u>sylbreak</u> (https://github.com/ye-kyaw-thu/sylbreak) က ကျွန်တော်တို့ မြန်မာစာကို unicode နဲ့သာ မုန်မုန်ကန်ကန်ရေးထားရင်၊ "Regular Expression တစ်ကြောင်းထဲနဲ့ လွယ်လွယ်ကူကူ ဖြတ်လို့ရကြောင်း" နောက်ပြီးတော့ အဲဒါက "တကယ်လည်း မြန်မာစာ NLP သုတေသနအလုပ်တွေအတွက် အသုံးဝင်ကြောင်း"၊ ဒါ့အပြင် "အခြေခံကျတဲ့ မြန်မာဝဏ္ဏ (syllable unit) အနေနဲ့လည်း ရပါလိ မ့်မယ်" ဆိုတဲ့ message ကိုပေးထားတာပဲ ဖြစ်ပါတယ်။ ကျွန်တော်ရဲ့ syllable unit တွေက မြန်မာစာအနေနဲ့ကြည့်ရင် ပြည့်စုံ မှန်ကန်တယ်လို့ မဆိုလိုပါဘူး။ ကိုယ်လုပ်မဲ့ အလုပ်၊ develop လုပ်နေတဲ့ application ပေါ်ကို မူတည်ပြီးတော့ လက်ရှိ ကျွန်တော် ပြင်ဆင်ပေးထားတဲ့ 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 သူများ အတွက် ကျွန်တော့်ရဲ့ <u>Tutorial (https://github.com/ye-kyaw-thu/Tutorials)</u> မှာ လေ့လာနိုင်ပါတယ်။