Digitizing Parliamentary Transcript

Nyan Lynn Myint
Ānanda Data
nyan@theananda.org

Why we do this?

- Transcripts are not searchable
- Not easily accessible by public

Pre Processing

- Scraping and Classifying the Transcripts
- Converting PDF Documents to Text
- Concatenation and Cleaning OCR Junks
- Normalization (Misspelled words, Word Segmentation, Irregular Whitespaces)

Structuring and Validation

- Building XML Skeleton
- Processing Documents
- Validating the Documents

Original Pdf

ဥက္ကဋ္ဌ။ ။ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာနက ဖြေကြားရန်ဖြစ်ပါတယ်။

ဖြေကြားချက်

ဒေါက်တာမြင့်ထွေး(ပြည်ထောင်စုဝန်ကြီး၊ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာန)။ ။ လေးစားအပ်ပါသော ပြည်သူ့လွှတ်တော်ဥက္ကဋ္ဌကြီးနှင့် ပြည်သူ့လွှတ်တော်ကိုယ်စားလှယ်များအားလုံး ကိုယ်၏ကျန်းမာခြင်း၊ စိတ်၏ချမ်းသာခြင်းနှင့် ပြည့်စုံပါစေကြောင်း ဦးစွာပထမဆုတောင်းအပ်ပါတယ်။ ကျွန်တော်ကတော့ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာန ပြည်ထောင်စုဝန်ကြီး ဒေါက်တာမြင့်ထွေး ဖြစ်ပါတယ်။ ဖလမ်းခရိုင်၊ ဖလမ်းမြို့ ရာပြည့်ဘောလုံးကွင်းတွင် မီတာ ၄၀၀ ပြေးလမ်းငါးခုပါသော စံချိန်မီဘောလုံးကွင်းအဖြစ်နဲ့ ဖလမ်းမြို့ကိုပြန်ကြည့်လိုက်တဲ့အခါမှာ ဘောလုံးကွင်းနှစ်ကွင်းရှိပါတယ်။ ရာပြည့်ဘောလုံးကွင်းနဲ့ ဟရန်ထီးယိုးဘောလုံးကွင်းရှိပါတယ်။ ရာပြည့်ဘောလုံးကွင်းကိုတော့ ၂၀၁၅-၂၀၁၆ ဘဏ္ဍာရေးနှစ်မှာ အဆင့်မြှင့်တင်ရေးလုပ်ငန်းအတွက် သိန်း ၆၀၀ ဖြင့် ဆောင်ရွက်ပေးပြီး ၂၀၁၇-၂၀၁၈ ဘဏ္ဍာရေးနှစ်မှာ သိန်း ၂၇၀ ဖြင့် ပွဲကြည့်စင်တည်ဆောက်ခြင်းလုပ်ငန်းကို ချင်းအလှကုမွဂၢီအမည်နဲ့ အခု လောလောဆယ်ဆောက်လုပ်လျက်ရှိပါတယ်။

Converted Text

ဥက္ကဋ္ဌ။ ။ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာနက ဖြေကြားရန်ဖြစ်ပါတယ်။ ဖြေကြားချက်
ဒေါက်တာမြင့်ထွေး (ပြည်ထောင်စုဝန်ကြီး၊ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာန)။ ။ လေးစားအပ်ပါသော ပြည်သူ့လွှတ်တော်ဥက္ကဋ္ဌကြီးနှင့် ပြည်သူ့လွှတ်တော်ကိုယ်စားလှယ်များအားလုံး ကိုယ်၏ကျန်းမာခြင်း၊ စိတ်၏ချမ်းသာခြင်းနှင့် ပြည့်စုံပါစေကြောင်း ဦးစွာပထမဆုတောင်းအပ်ပါတယ်။ ကျွန်တော်ကတော့ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာန ပြည်ထောင်စုဝန်ကြီးဒေါက်တာမြင့်ထွေး ဖြစ်ပါတယ်။ ဖလမ်းခရိုင်၊ ဖလမ်းမြို့ ရာပြည့်ဘောလုံးကွင်းတွင် မီတာ ၄၀၀ ပြေးလမ်းငါးခုပါသော စံချိန်မီဘောလုံးကွင်းအဖြစ်နဲ့ ဖလမ်းမြို့ကိုပြန်ကြည့်လိုက်တဲ့အခါမှာ ဘောလုံးကွင်းနှစ်ကွင်းရှိပါတယ်။ ရာပြည့်ဘောလုံးကွင်းနဲ့ ဟရန်ထီးယိုးဘောလုံးကွင်းရှိပါတယ်။ ရာပြည့်ဘောလုံးကွင်းကိုတော့ ၂၀၁၅၂၀၁၆ ဘဏ္ဍာရေးနှစ်မှာ အဆင့်မြှင့်တင်ရေးလုပ်ငန်းအတွက် သိန်း ၆၀၀ ဖြင့် ဆောင်ရွက်ပေးပြီး ၂၀၁၇–၂၀၁၈ ဘဏ္ဍာရေးနှစ်မှာ သိန်း ၂၇၀ ဖြင့် ပွဲကြည့်စင်တည်ဆောက်ခြင်းလုပ်ငန်းကိုချင်းအလှ ကုမ္ပဏီအမည်နဲ့ အခု လောလောဆယ်ဆောက်လုပ်လျက်ရှိပါတယ်။
မူကြမ်း_____

Output XML

```
<speech type='speaker'>
ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာနက ဖြေကြားရန်ဖြစ်ပါတယ်။
</speech>
<speech><agency id='20'>ဒေါက်တာမြင့်ထွေး(ပြည်ထောင်စုဝန်ကြီး၊ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာန)</agency>
လေးစားအပ်ပါသော ပြည်သူ့လွှတ်တော်ဥက္ကဋ္ဌကြီးနှင့် ပြည်သူ့လွှတ်တော်ကိုယ်စားလှယ်များအားလုံး ကိုယ်၏ကျန်းမာခြင်း၊ စိတ်၏ချမ်းသာခြင်းနှင့်
ပြည့်စုံပါစေကြောင်း ဦးစွာပထမဆုတောင်းအပ်ပါတယ်။ ကျွန်တော်ကတော့ ကျန်းမာရေးနှင့် အားကစားဝန်ကြီးဌာန ပြည်ထောင်စုဝန်ကြီး
ဒေါက်တာမြင့်ထွေး ဖြစ်ပါတယ်။ ဖလမ်းခရိုင်၊ ဖလမ်းမြို့ ရာပြည့်ဘောလုံးကွင်းတွင် မီတာ ၄၀၀ ပြေးလမ်းငါးခုပါသော
စံချိန်မီဘောလုံးကွင်းအဖြစ်နဲ့ ဖလမ်းမြို့ကိုပြန်ကြည့်လိုက်တဲ့အခါမှာ ဘောလုံးကွင်းနှစ်ကွင်းရှိပါတယ်။ ရာပြည့်ဘောလုံးကွင်းနဲ့
ဟရန်ထီးယိုးဘောလုံးကွင်းရှိပါတယ်။ ရာပြည့်ဘောလုံးကွင်းကိုတော့ ၂၀၁၅-၂၀၁၆ ဘဏ္ဍာရေးနှစ်မှာ အဆင့်မြှင့်တင်ရေးလုပ်ငန်းအတွက်
သိန်း ၆၀၀ ဖြင့် ဆောင်ရွက်ပေးပြီး ၂၀၁၇-၂၀၁၈ ဘဏ္ဍာရေးနှစ်မှာ သိန်း ၂၇၀ ဖြင့် ပွဲကြည့်စင်တည်ဆောက်ခြင်းလုပ်ငန်းကို
ချင်းအလှ ကုမ္ပဏီအမည်နဲ့ အခုလောလောဆယ်ဆောက်လုပ်လျက်ရှိပါတယ်။
```

Thank You