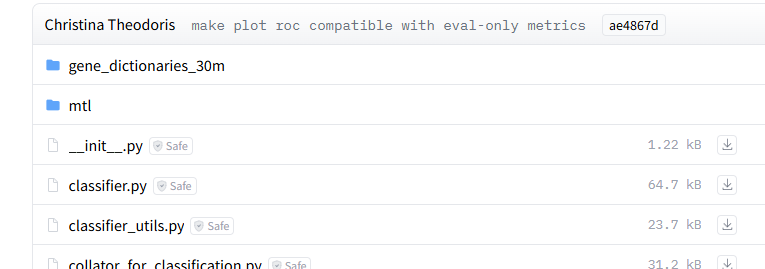
**3 FILES To CHECK BEFORE TOKENIZING**









1. Input file should be a directory containing .loom or h5ad file -containing raw counts from single cell RNA data
2. Genes should be labelled as Ensemble ID - .loom row attribute –**"ensembl\_id"**
3. Cells should be labelled with their row counts **(n\_counts).**
4. Cell metadata is not required – but cell attributes can be passed, eg if original data is having attributes organ\_major and cell\_type and one like to retain these attributes, one can pass like

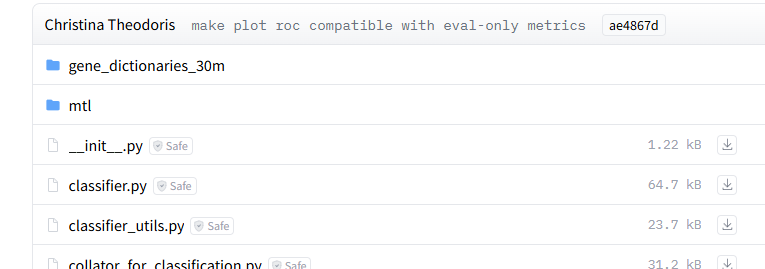
**{"cell\_type": "cell\_type", "organ\_major": "organ"}.**

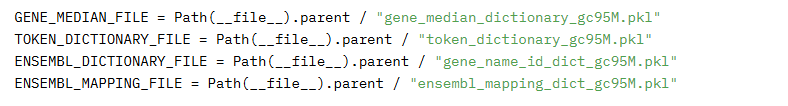
1. **"filter\_pass":**  all cells with 1 will be tokenized whereas others will be excluded.
2. 95 Million series **special\_token** =True

30 Miilion series **special\_token** = False

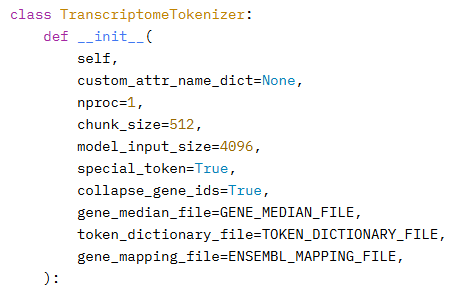
1. Use right gene median dictionary and token dictionary in .init file.











1. Change model input size 4096 or 2048
2. special\_token= True for 95M series

special\_token= False for 30M series

