Add chapter about data preprocessing

OWT -> Null bytes, empty lines etc. need to be removed

D2V image does not use a cls token -> Can’t be used for me then. Does is influence the kd target creation since we take the first time step, which is usually the special token?

Nevermind, it does use it (whenever “num\_extra\_tokens” >0 )

->

<https://github.com/facebookresearch/fairseq/blob/bedb259bf34a9fc22073c13a1cee23192fa70ef3/examples/data2vec/models/modalities/base.py>

Audio has no special token!

Only for text it has special token.

For text, only prepend token is used -> No eos token -> I can’t use that!

What is the strategy to aggregate the time steps? Mean, special token (if available), or both?

Availabilty for Modalities:

Text: BOS token (at beginning)

Audio: None -> Only mean possible

Image: CLS token

You are dependent on it, because this is how the modality encoders are trained, and they are now frozen!

Strategies for time step reduction in kd pretrain:

Mean target, Mean pred;

Mean target, special token pred (if available, else fallback to 1.);

Special token target (if available), mean pred;

Special token target(if available), special token pred (if available, else fallback to 1.);