Clarisse Taboy
HW2 — Language Technology
October 26, 2021

WER: 32

Challenges:
1. It took me a long time to figure out where I could find the arguments I needed for the training portion, since not all the parameters provided appeared in the fairseq docs. But eventually, after running the command with my best guesses, I was reminded by the error messages that I could just use the '—h' argument on fairseq-train to see exactly what I needed.

2. For the last parameter in fairseq-train that says 'don't store epoch checkpoints', it took me a few minutes to decide whether I should use '——no-save' or '——no-epoch checkpoints' based on their respective descriptions in the fairseq docs. I found them to be a little vague, since neither description really specified if the checkpoints they were referring to were epoch checkpoints or not. At the time, I also wasn't sure what 'epochs' were, so I watched a video or two on YouTube that explained it. In the end, I went with '——no-epoch-checkpoints' based on its phrase similarity to the hint, and because it would produce two large files for the last and best checkpoints as per the instructions.

3. Looking at the generated predictions.txt, I saw that a few of the target Icelandic words had '<unk>' in their transcription, which to my understanding corresponds to the ——thresholdtgt parameter from the model training step. My guess is that this bumped up my WER since I was comparing whether the target words were equal to the predicted words. I'm not sure if this is a big deal, or if there are ways to approach cases when words have '<unk>'.