IS 784 Deep Learning for Text Analysis

Assignment 1 (Deadline 25 April 2021)

1. Shared notebook

A notebook file given with this file contains some empty code blocks marked as “Fill Here”. Fill them to make the network work.

2. Improving the Preprocessing

Create a copy of the working network at part 1;

Improve preprocessing steps by editing “tokenize”, “preprocessing” and “postprocessing” variables of the Field class by;

a) Only stemming
b) Only lemmatization
c) Lemmatization and steps like negation handling, removing stop word (and anything you think it may be useful in this dataset)

Then compare the performances of the models with each other and results obtained at part 1 (the initial model). Report their performance scores in detail (choose at least two different relevant metrics) .

Note: The order of operations for Field is as follows tokenize->preprocessing->numericizing->postprocessing->tensor creation .

You can submit one notebook, but don’t forget to implement all 3 preprocessing steps as different functions.

3. Improving the Network

Create a copy of one of the networks from the previous questions (preferably with highest performance).

a) Try different embedding sizes.
b) Reduce the vocabulary size by truncating less frequent words.
c) Use pretrained word vectors with and without freezing embedding layer. (These vectors may not handle negation and mark all negative words as <unk>)
d) Increase and decrease the depth of the network.

How do these changes affect the performance of the network? (Consider metrics like accuracy, training time so on…) Discuss your results.