CS288 HW5: Retrieval and Question Answering

Nicholas Tomlin and Dan Klein

Due: 24 April 2022, 11:59PM PST

Overview

This homework will be focused on building a retrieve-and-read question answering system. Once again, this homework has a large amount of stencil code, which we recommend reading carefully to make sure you're understanding the material. The core material of this homework involves building unsupervised retrievers with TF-IDF and BM25 and a neural reader model based on DistilBERT. For your report, you'll also have the opportunity to experiment with a pre-trained dense passage retrieval model. The latter part of this homework is adapted from MIT 6.864 by Jacob Andreas, Jim Glass, and Yoon Kim.

Background Reading

Feel free to consult with the following resources before beginning this assignment:

- Question Answering: `https://web.stanford.edu/~jurafsky/slp3/23.pdf`
- Dense Passage Retrieval: `https://arxiv.org/abs/2004.04906`

Assignment

Notebook: `https://www.kaggle.com/nickatomlin/cs288-hw5-public`

- Complete the the notebook and save the generated files
- Write a report about your exploratory experiments with the dense passage retrieval model

Submission to Gradescope

Please submit the assignment to: `https://www.gradescope.com/courses/361823/` (code: 4PBP57)

When you upload your submission to the Gradescope assignment, you should get immediate feedback that confirms your submission was processed correctly. Be sure to check this, as an incorrectly formatted submission could cause the autograder to fail. In particular, please make sure that you're zipping your files and not the folder containing them. Note that Gradesope will allow you to submit multiple times before the deadline, and we will use the latest submission for grading. Make sure you have the following files (with correct names and extensions):

- `bm25_predictions.json`
- `retrieve_and_read.json`
- `hw5.ipynb`
- `report.pdf`

Please note that certain cells in `hw5.ipynb` should not be modified and are marked accordingly; modifying these cells could affect your grade, because they are related to data splits and evaluation.