CSS Bootcamp
Neural Analysis of Text Data
Day 3: Tokenizing, Padding, and Batching

Xingyuan Zhao

Sept. 14 2022

1 Sentiment Analysis

Find the most positive and negative sentence in the paragraph below.

"Eighteen years have gone by, and still I can bring back every detail of that day in the meadow. Washed clean of summer's dust by days of gentle rain, the mountains wore a deep, brilliant green. The October breeze set white fronds of head-tall grasses swaying. One long streak of cloud hung pasted across a dome of frozen blue. It almost hurt to look at that faroff sky. A puff of wind swept across the meadow and through her hair before it slipped into the woods to rustle branches and send back snatches of distant barking-a hazy sound that seemed to reach us from the doorway to another world. We heard no other sounds. We met no other people. We saw only two bright, red birds leap startled from the center of the meadow and dart into the woods. As we ambled along, Naoko spoke to me of wells. Memory is a funny thing. When I was in the scene, I hardly paid it any mind. I never stopped to think of it as something that would make a lasting impression, certainly never imagined that eighteen years later I would recall it in such detail. I didn't give a damn about the scenery that day. I was thinking about myself. I was thinking about the beautiful girl walking next to me. I was thinking about the two of us together, and then about myself again. It was the age, that time of life when every sight, every feeling, every thought came back, like a boomerang, to me. And worse, I was in love. Love with complications. Scenery was the last thing on my mind. Now, though, that meadow scene is the first thing that comes back to me. The smell of the grass, the faint chill of the wind, the line of the hills, the barking of a dog: these are the first things, and they come with absolute clarity. I feel as if I can reach out and trace them with a fingertip. And yet, as clear as the scene may be, no one is in it. No one. Naoko is not there, and neither am I. Where could we have disappeared to? How could such a thing have happened? Everything that seemed so important back then-Naoko, and the self I was then, and the world I had then: where could they have all gone? It's true, I can't even bring back Naoko's face-not right away, at least. All I'm left holding is a background, sheer scenery, with no people up front."

1.1 Finish the same task as we did on Day 1, instead of using pipeline, let's do it from tokenizer to model, and analyze the vanilla output from scratch.

1.2 Finish the same task as above, instead of using the tokenizer directly, do tokenize, convert tokens into token_ids, add start and end token, do padding and generate attention mask by your own.

1.3 Finish the same task with batch size of 4