question @117 30 views

HW3 flow question

I wonder if my understanding of HW3 flow is correct: I need to crawl 40000 web pages, and then index them to my elastic search, and then I need to save these 40000 documents to my local computer in the same format as the AP89 corpus. And my teammates should do the same. When we are all done, I need to copy their 80000 documents to my local computer, then parse them to get the DOCNO, text and index them to my elastic search. Finally, my elastic search will have 120000 documents, the “Merge” part is done.

hw3

Updated 5 days ago by Anonymous Beaker

the students' answer, *where students collectively construct a single answer*

Do we need to first save 40000 documents to AP89 format and then start indexing them to ES?

Updated 4 days ago by Anonymous Scale

the instructors' answer, *where instructors collectively construct a single answer*

The idea is to crawl 40k webpages, format them based on the AP89 data structure
Your team mates will do the same.
Then you have to add the docs to elastic search, you can make use of the HW1 code for this. There are ways to doing this, either you can get together and merge in a parallel motion, else you can do one by one using elastic search cloud.

Updated 4 days ago by Raaghavv Devgon

followup discussions *for lingering questions and comments*