note @122 14 views

Elastic Cloud Merging

Folks can use Elastic Cloud for performing merging.

Here is the below link

https://www.elastic.co/cloud/

It offers a 14 day trial w/o credit card, so I'd urge to use this wisely, I'd recommend one person of the team signing up for this and sharing their credentials.

How to use it?

Once you sign up, you will get a link through which you can access the es instance on cloud, below is the one I had (password doesn't work anymore :)), you can just re use your homework 1 code, and switch it from localhost to the below config.

```
host='https://elastic:KTPRMHa5BA5fjrh8zwqqIZOQ@0bb7e3d7643c42ae85d3d8ad417593f3.us-east-1.aws.found.io:9243'
es = Elasticsearch([host],timeout=3000)
print(es.ping())
```

run code snippet

if es.ping results in anything apart from an error, you are connected , you can also try to index some data to see if it is working.

For merging, I'd recommend having the same.'id' creation, remember the docids you created for HW1, well you have to do something like that for this.

The best way to do it is by using url for an id, the same canonicalizer is to be used across the team, so this makes it easier for you to know if you and your team members have accessed the same link!

You can also create a MD5 hash out of the url, if you want to make it cleaner, it is up to you.

How to do merge?

- Do it turn by turn: Suppose there is a team of A,B,C.

A decides to index the data first

The index fields are as follows (you have to store the below fields in AP89 format):

-id

-body/text

-inlinks

-outlinks.

-author (your name)

You can modify the hw1 code/reuse it to make this happen.

Now A has index their 40k docs,

It's B's turn now,

B is likely to have crawled some links which A's already indexed. So how should B index?.

B should first query ES, see if a particular id is present in the index. (search api maybe helps you to do that?)

if there is, it should retrieve the inlinks and outllinks, if B has some inlinks/outlinks which haven't been indexed it should append and re index that, and also change the name of the author as AB, as it has been merged using the contents of A and B. This will also help you to see if the merge has worked correctly.

If B has a link A hasn't crawled, it can just index the data like A did.

The index fields are as follows (you have to store the below fields in AP89 format):

-id

-body/text

-inlinks

-outlinks.

-author (your name)

C will also follow the same procedure as B and then you have a merged index.

If you have any questions/concerns, feel free to visit OH.

Good luck
#pin

hw3

Updated 6 hours ago by Raaghavv Devgon

followup discussions *for lingering questions and comments*