Logger of 21 Nov 2017

Questions

- The fonction 'is404erreur' in the error404detector.py, there is a faut the line 113 and line 121: we should return "False", if we find the match, and the line 123: the "False" must be replaced by "True", if we have not found the match. PS: Is this a simple detector, cause I saw some cases which can not be detected?
- When do you prefer to finish this task? I still need to do some NLP and apply the trainning methods. And, I think that we still need to review one or two times this problems.
- For the training mathods, do you have some preferences or we'll do stacking or just pick up one radomly?

Overview of what I have done

- I added my ssh-key in ~/.ssh/ on the server and get all the configuration things done.
- Preprocessing: I have made the first trial to analyse this problem. I tried a lot and I insist a lot on the point of NLP, but I have had some problems with cleaning these htmls, so I spent a lot of time in trying to wash it and searching on the Internet if there are some packages which can cover this dirty job. And finally, I use a package called "Beautiful Soup".
- Analysis: I did some research over the Length of Text, which offered me some results interesting.

My thoughts

- Cause what we got is all these htmls gotten by spiders(I guessed), so I treat this problem as a problem a NLP problem which I do not have encoutered before.

Work left

- The part of NLP analysis
- Application of trainning methods
- Perhaps several reviews (discussion) on this problem