

## QTM385 HW3 REPORT

### Question1:

My method `change_email_dict` is able to detect the author who uses [first.last@school.edu](mailto:first.last@school.edu) as well as [firstname.lastname@school.edu](mailto:firstname.lastname@school.edu) and so on. Most of the mismatches are because whether the author didn't write an email or he used the format described previously. I first parse out the domain and add the domain after the author's name for each author using this format. By doing this can reduce most of the mismatches. For example for ID 'C10-1012', the tsv file now can show all four emails correctly. However, for those mismatches who have both the correct email and the formatted email this method will recount for those correct one. It is hard to handle all situations.

### Question2:

Method `weight_count` counts how many different domains appear for each ID and for how many times, but I have some problems adding the result to the fourth column of `email_map.tsv`.