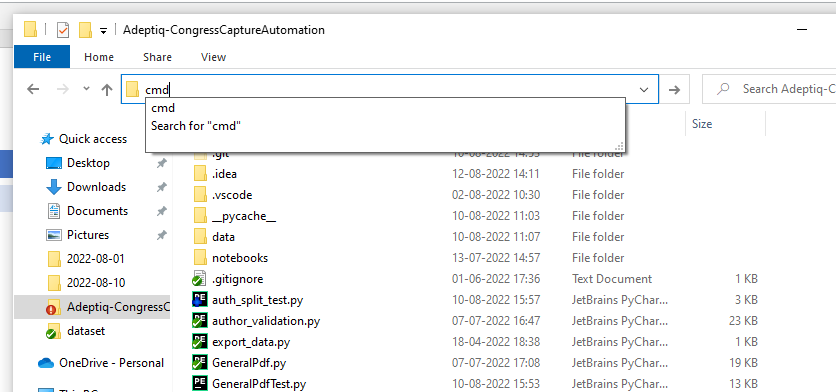
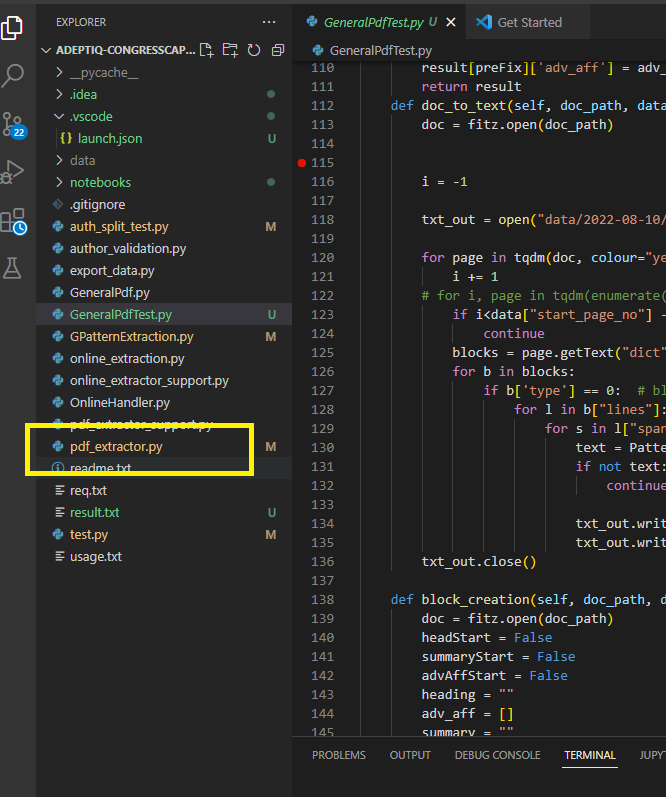
# Congress usage

## running

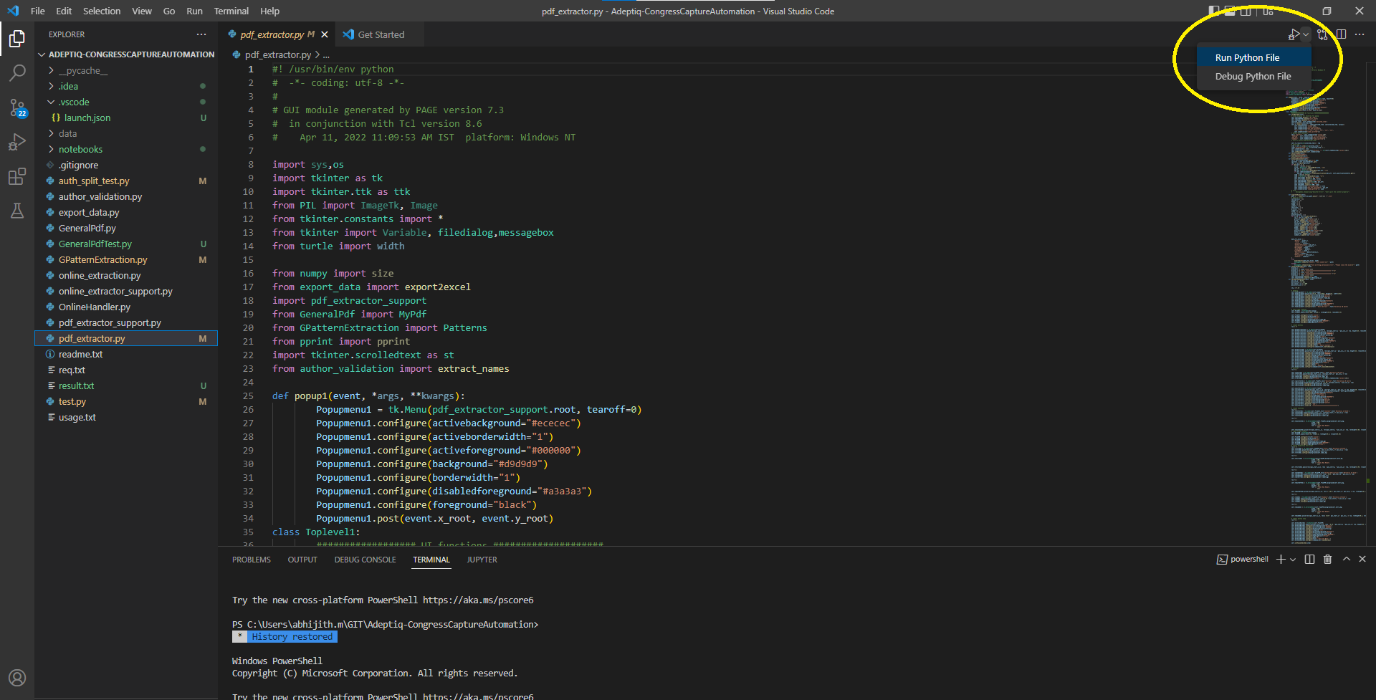
* Go to project location and type “cmd” in address bar



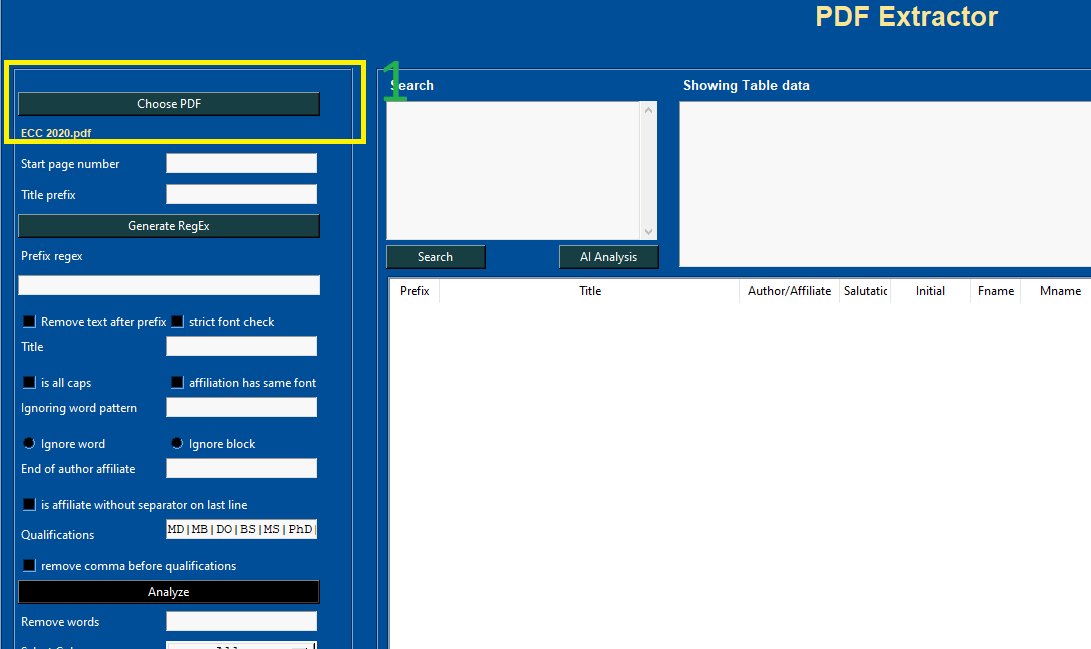
* On the command prompt type
* code .
* press enter
* in vs code open “pdf\_extractor.py”

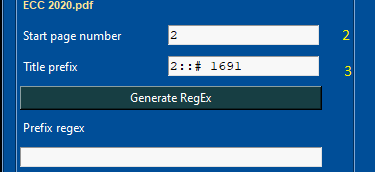


* then press the start button

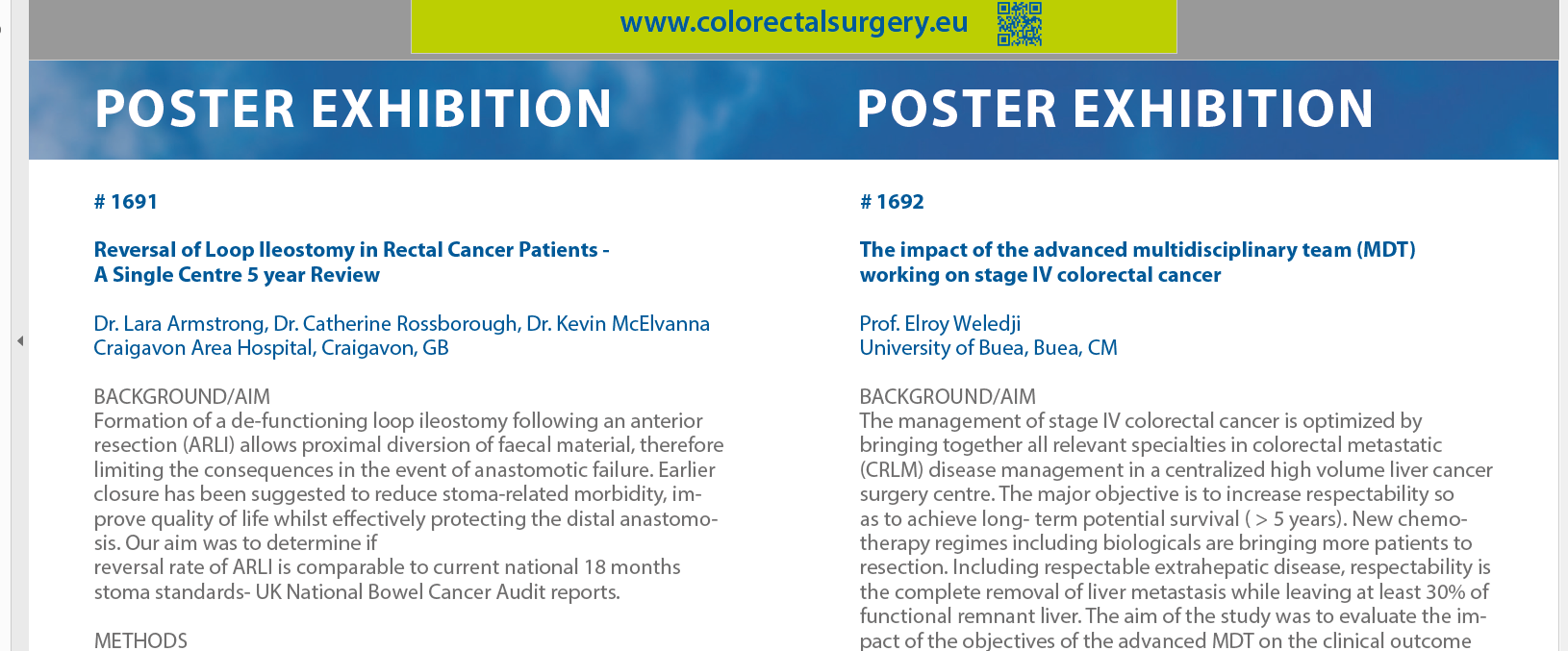


## Pdf\_extractor

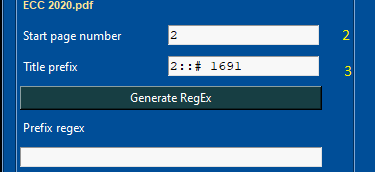




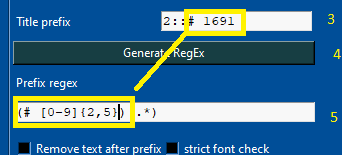
2 – start page number: the page which starting abstract

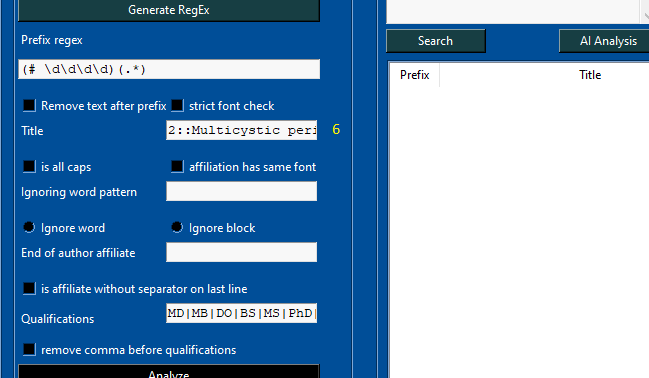


In tool



3 – title prefix: page number :: prefix



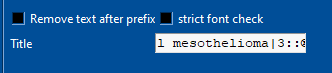
* press generate regEx – this will create a regular expression for prefix
* \d – means a number
* So if we face any accuracy issue, we can replace this regEx with our won logic
* 

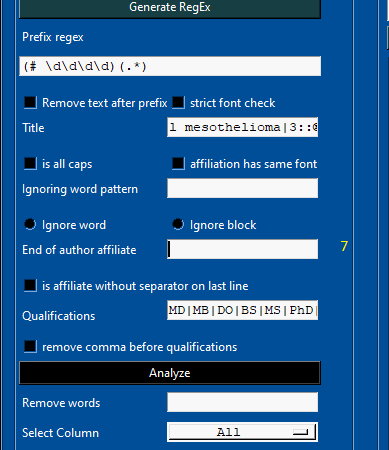
6 – page number :: abstract title

This is for detecting font of heading we can add multiple font by separating a | symbol

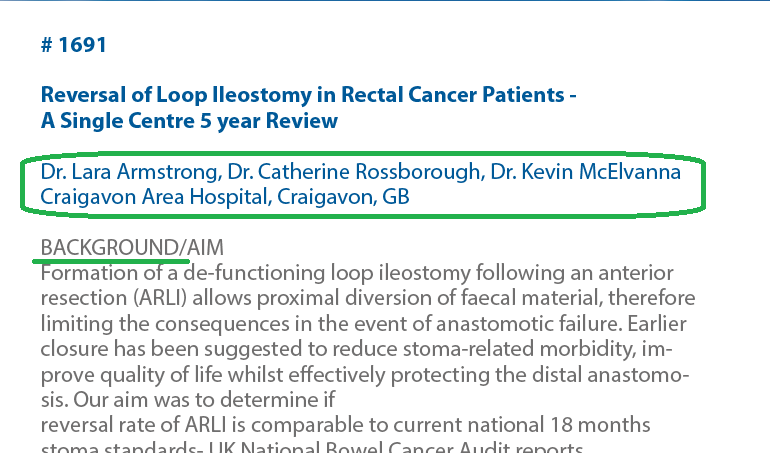
For example the symbol **®** has a different font



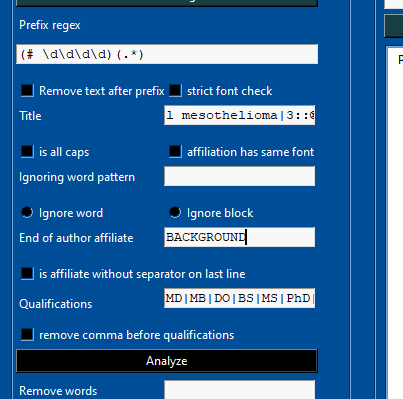




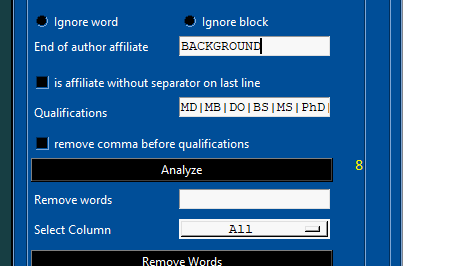
7 – end of author/ affiliation:



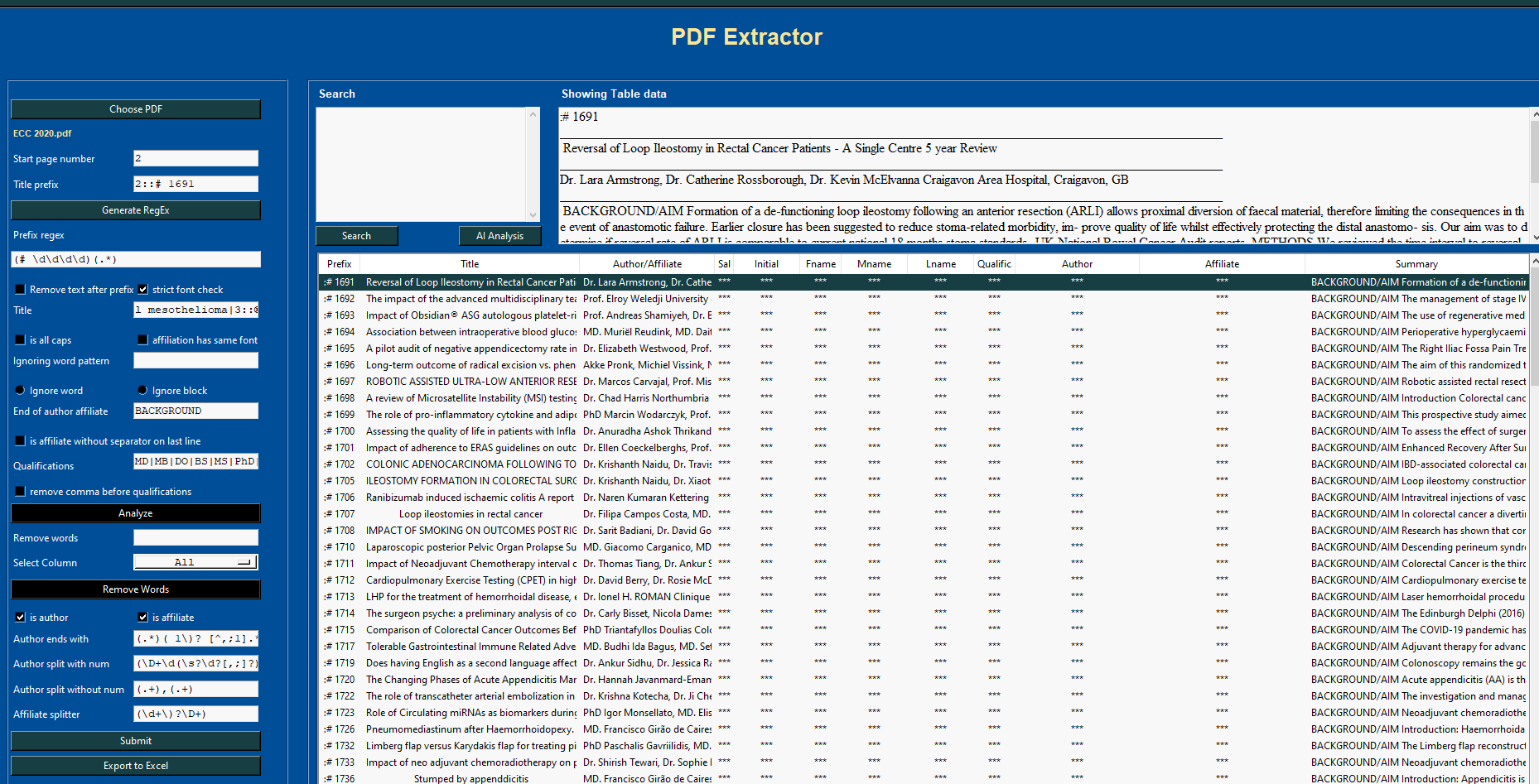
The author/ affiliation is ending with the word “BACKGROUND”

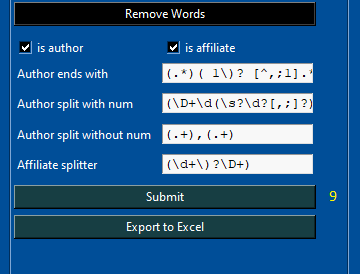


* We can add multiple author/affiliation by using | symbol



8 – this will generate a table based on our input

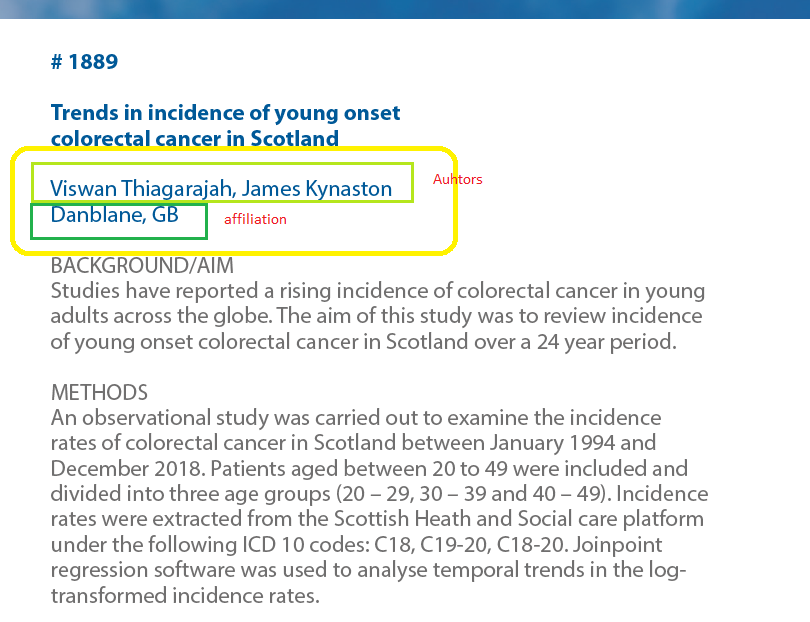


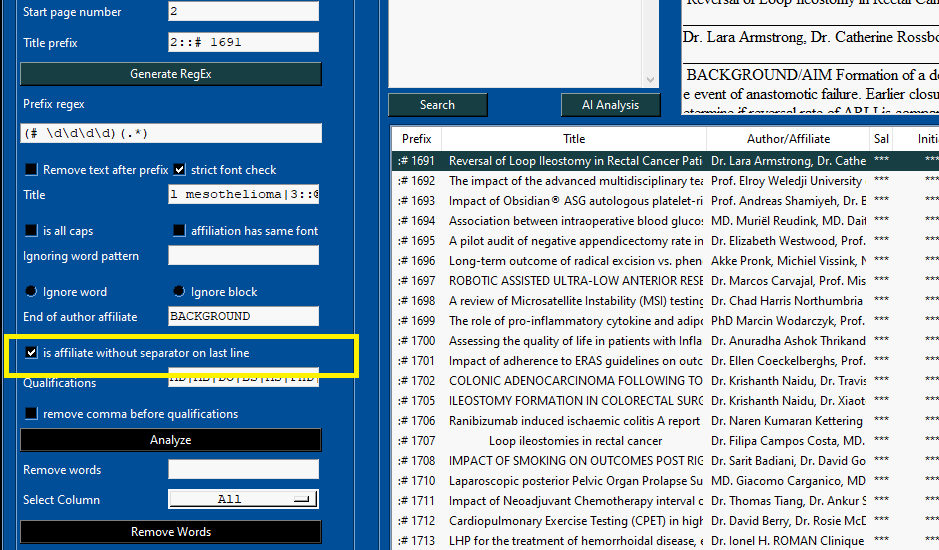


9 – this submit button will split author and affiliates

90% of cases we do not need to update this input field

Some pdf their may be a case, the affiliation is on last line with no separator(like ;,.) so we need to specify the affiliation is on last line with out separator





After splitting author / affiliate. If we want to go back to the previous session

Type 111111111 in remove words the press “remove words” button

