In Second problem I have first defined a function which takes all the links that is stored in a list “urls”. From all the urls html tags are extracted in which everything is cleaned and only the words are being extracted.

All the cleaned words are saved in a list which removes the dot symbols

Get\_word\_count function checks each word if it is repeating and increases the count

I have parsed the main webpage and extracted all the links in which if the link is not working it just ignores and also ppt extension links will also be ignored and pdf links are handled separately as

I assumed all the Pdf files are saved in the disk and text is extracted and it is appended to the text that came from all the other links

Finally, all the words are being counted and how many documents the words in a main page occurs in other documents

Occurrences is a dictionary that have the words of main page that occur in the other document

Output is stored in two text files