Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests

Now-a-days almost all users are using web pages to get various information such as news, sports, technology etc but all web pages will use noise data such as images, video clips or advertisement which makes difficult for the users to get interested information. To remove noise data all existing technologies were using static web matching pattern such as the main page look and feel will be match with rest of the screen and if not match then it will remove unmatched data from the web pages to show only interested data to the user. This static technique will not work if web pages look and feel changes dynamically.

To overcome from above issue author is proposing Noise Web Data Learning (NWDL) technique, in this technique server will maintain log for each user access page and will be called as web log dataset. This dataset will have information such as User\_id, access\_page, date\_time, URL. By analyzing such log data we can identify user interested pages in dynamic or static web pages. User interested pages can be found by seeing frequency of web page access by a single user and total time spend on each page.

If user spend more time and access this page more than 2 times then we can consider that user is interested in that page. If user spend less time on seeing that page and visiting that page very rarely then it will consider as uninterested page and will be called as noise page.

Dataset Example

User\_id access\_page date\_time

1 abcd.html 2019-01-22 11:00:12

2 xyz.html 2019-01-22 12:18:23

1 abcd.html 2019-01-22 11:05:18

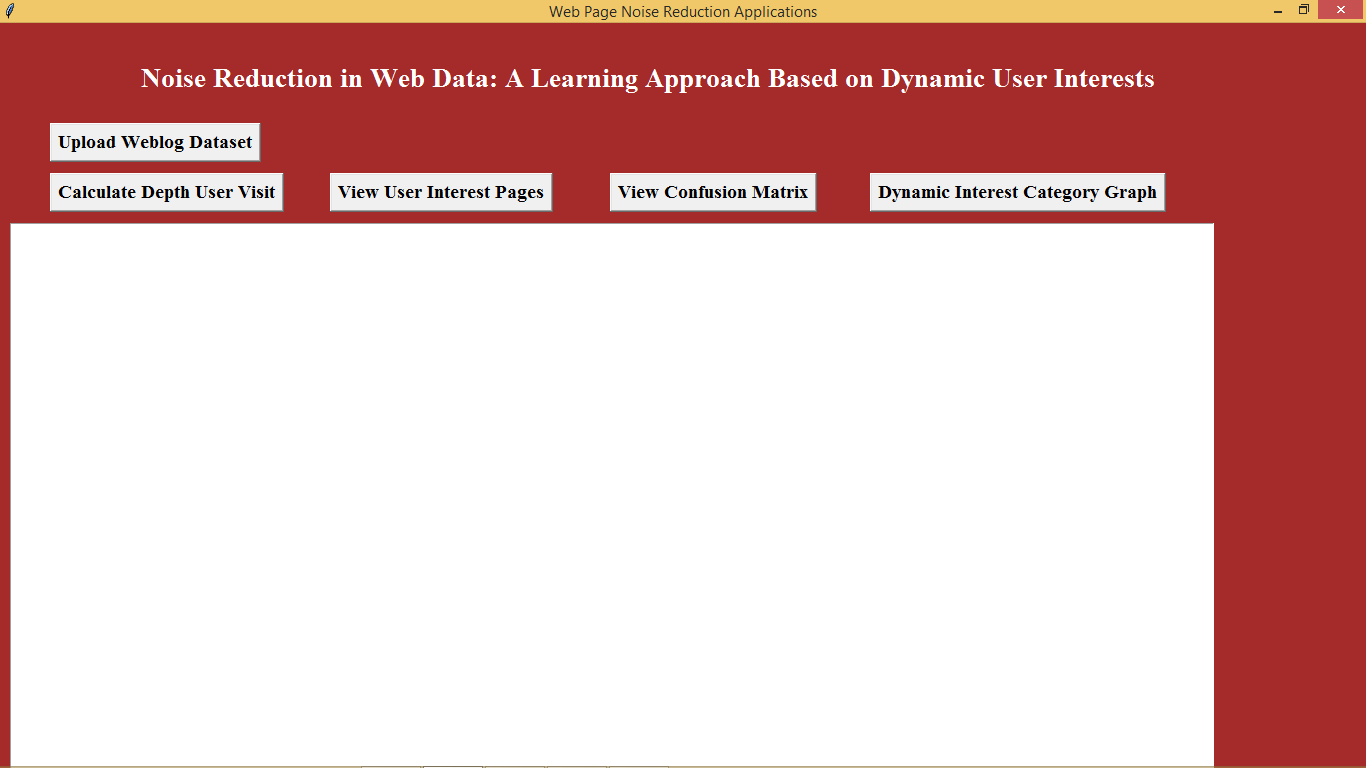
1 abcd.html 2019-01-22 11:06:12

1 xyz.html 2019-01-22 12:22:23

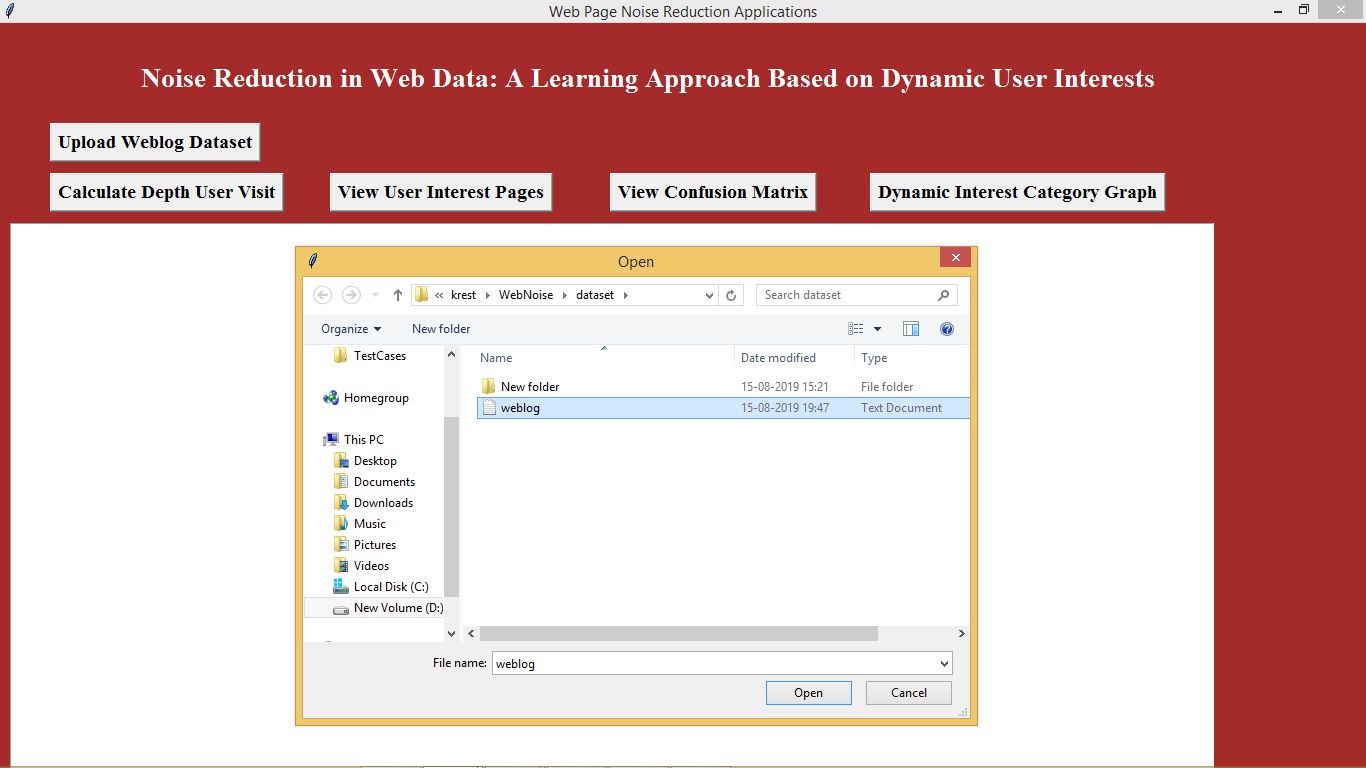
From above web log dataset we can easily says that user 1 accessing abcd.html more no of time and its frequency is 3 and he spend almost 6 minutes on that page (spend time will be calculated from first same page visit to till last same page visit) and this abcd.html will be consider as interested page from user 1 and xyz.html is rarely access by that user and will be consider as noise page and will not recommend to user.

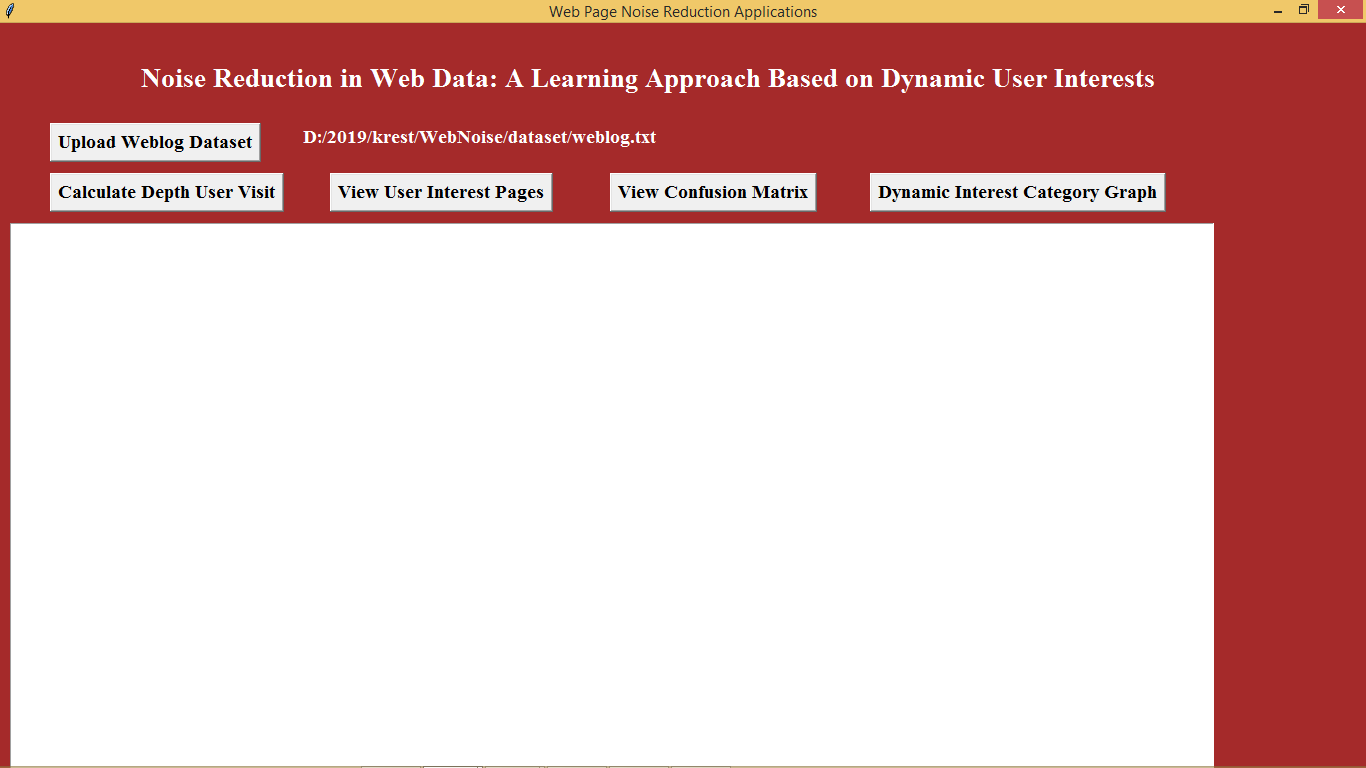
To perform experiment author has used WEBLOG dataset and i am also using same dataset and this dataset is available inside ‘dataset’ folder.

Screen shots

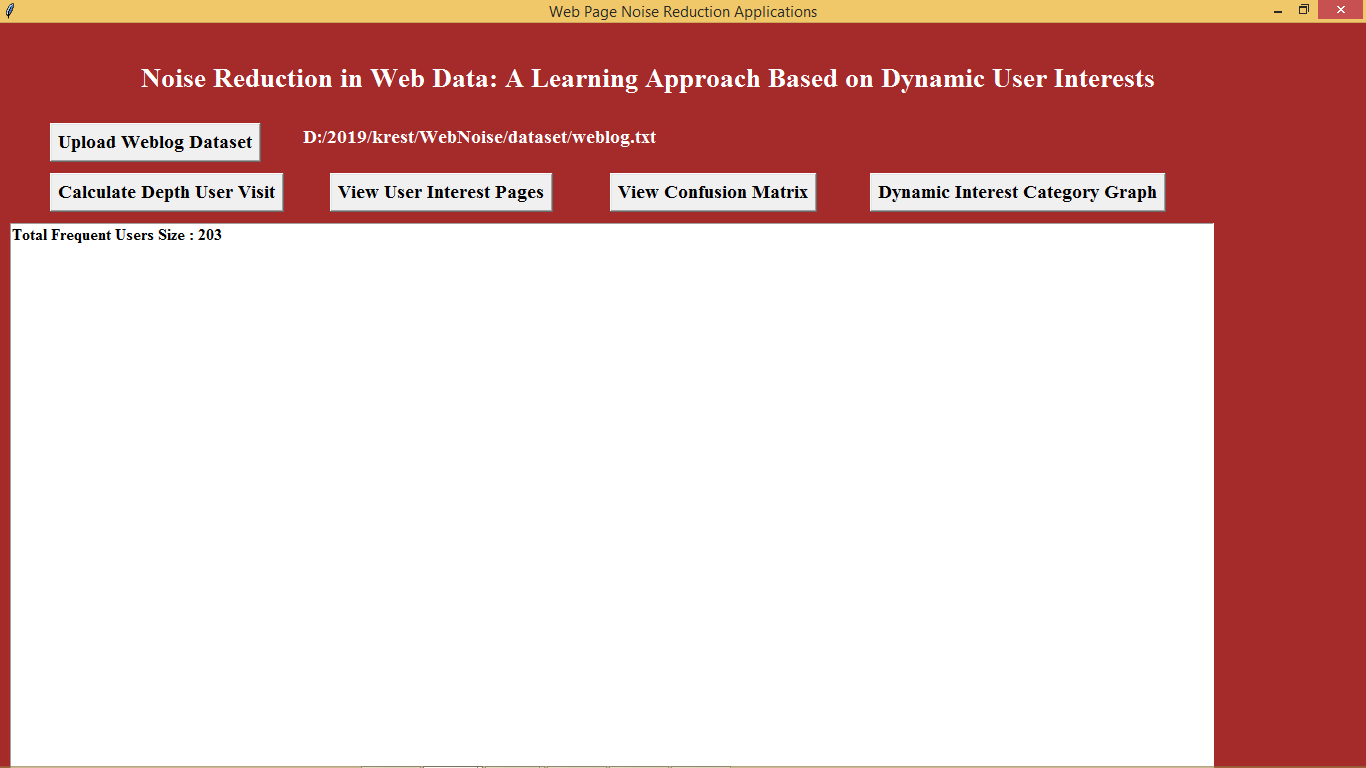


In above screen click on ‘Upload Weblog Dataset’ button to upload dataset

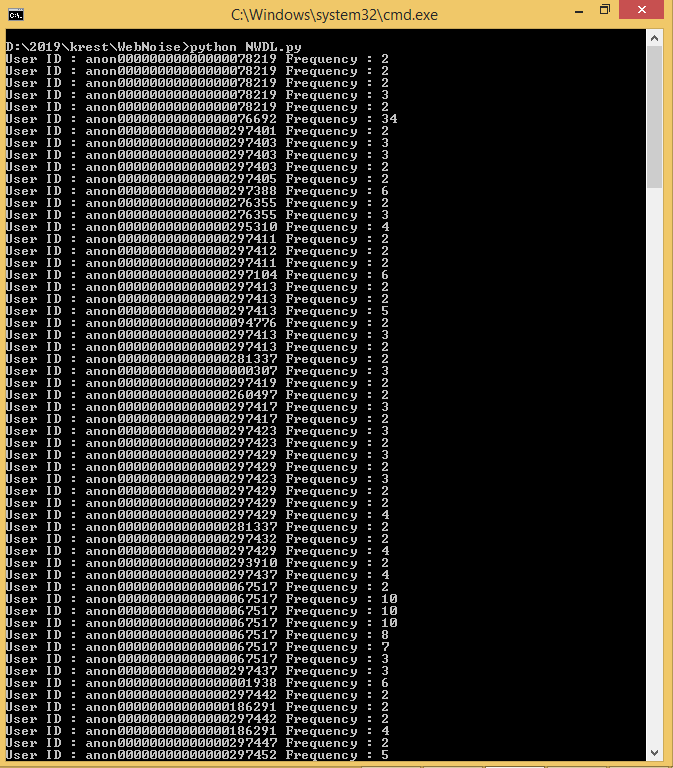




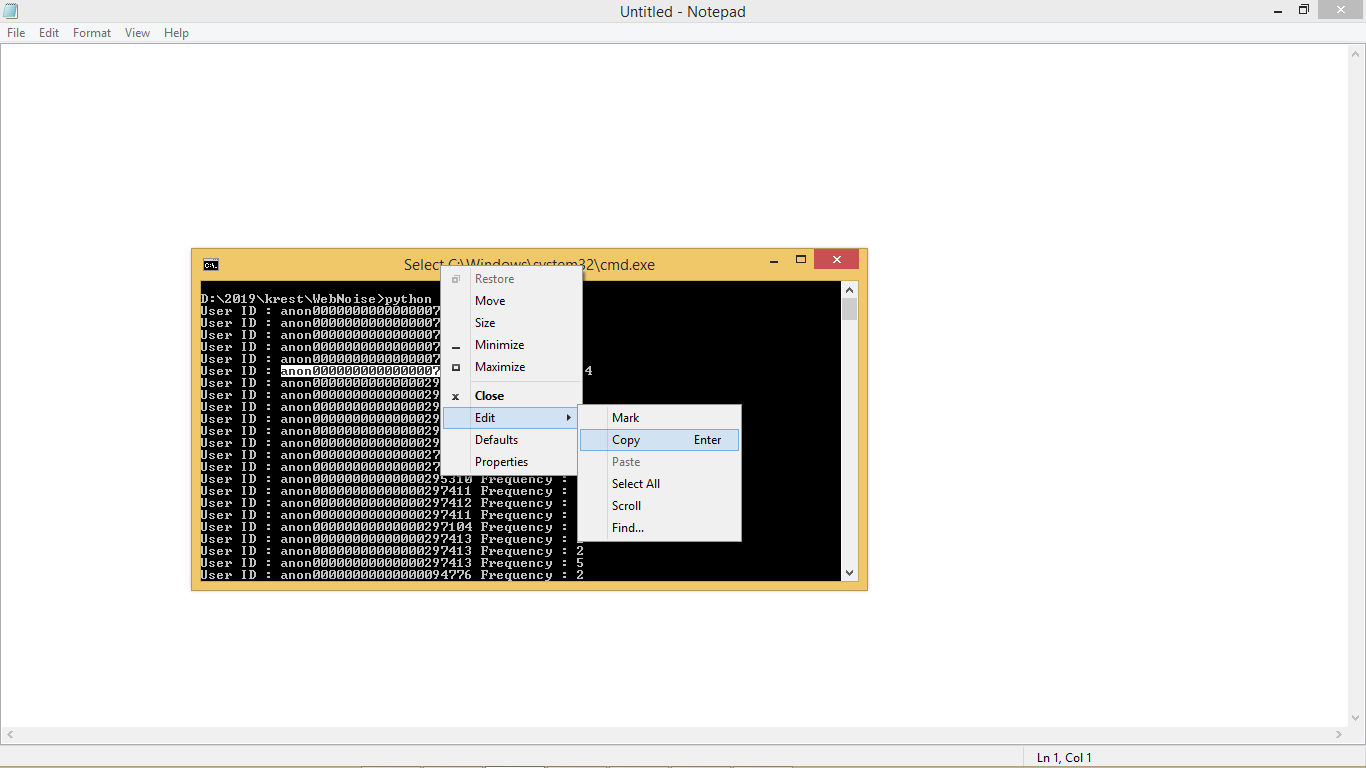
After dataset upload click on ‘Calculate Depth User Visit’ button to calculate frequency and weight of each page visit by single users



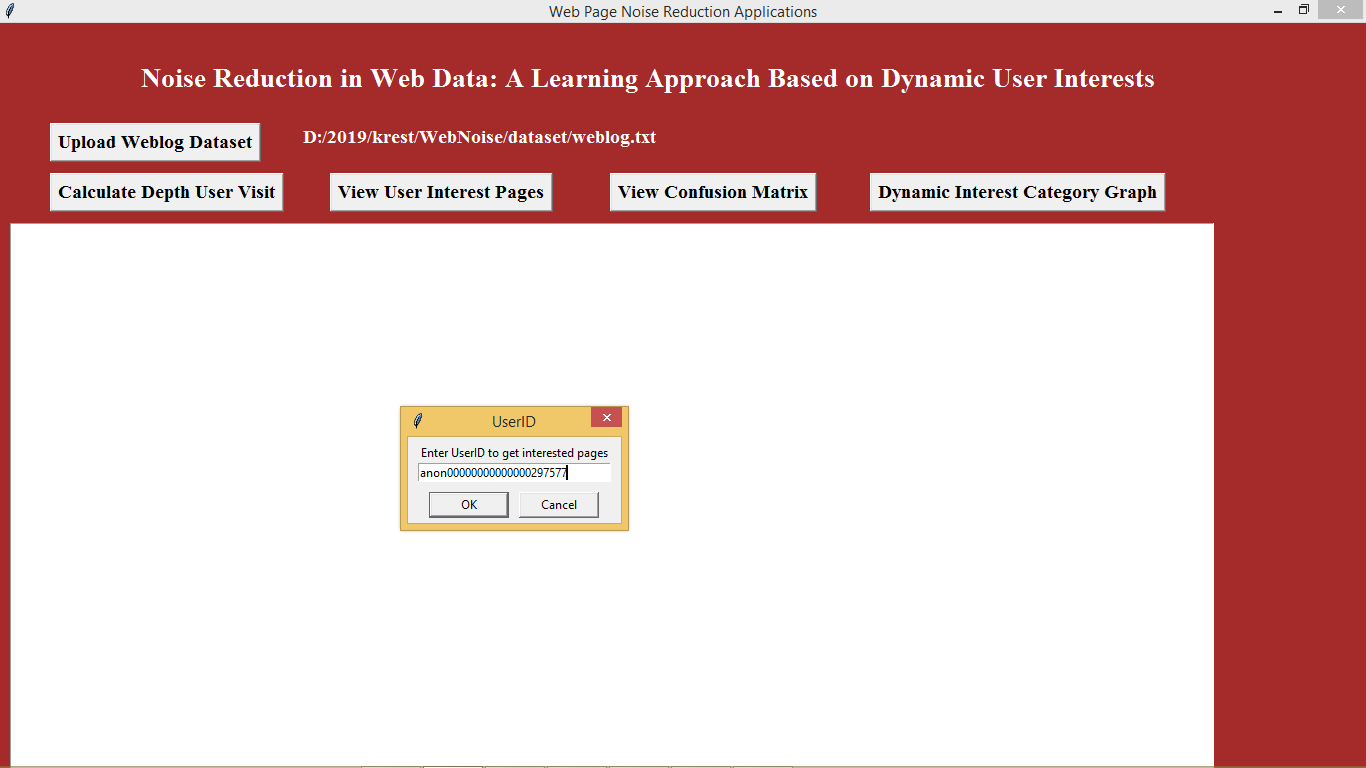
In above screen we can see total 203 web pages which access more frequently. To see frequency of each access page see command prompt console. See below screen



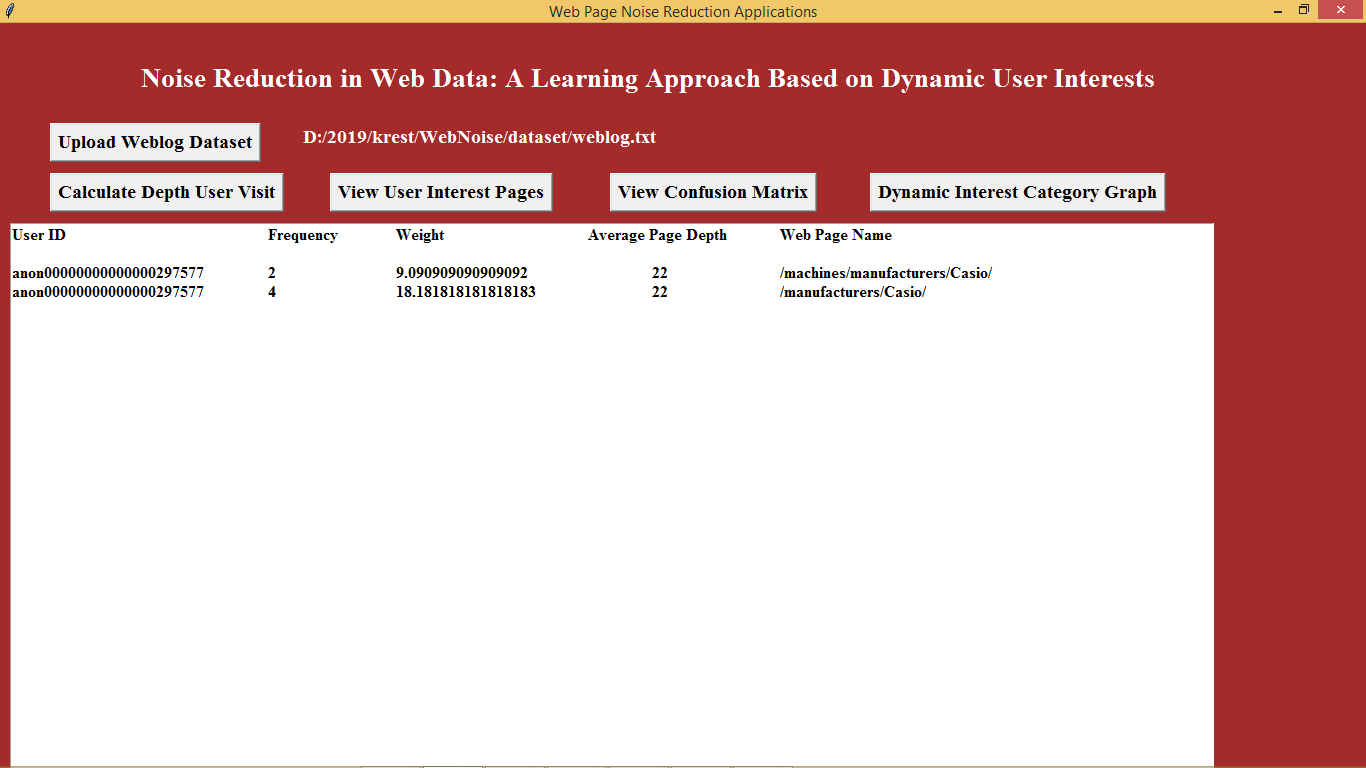
In above screen we can see user id and frequency of each page access by them, if u wants to see page name and weight details then copy one user id from black console. See below screen

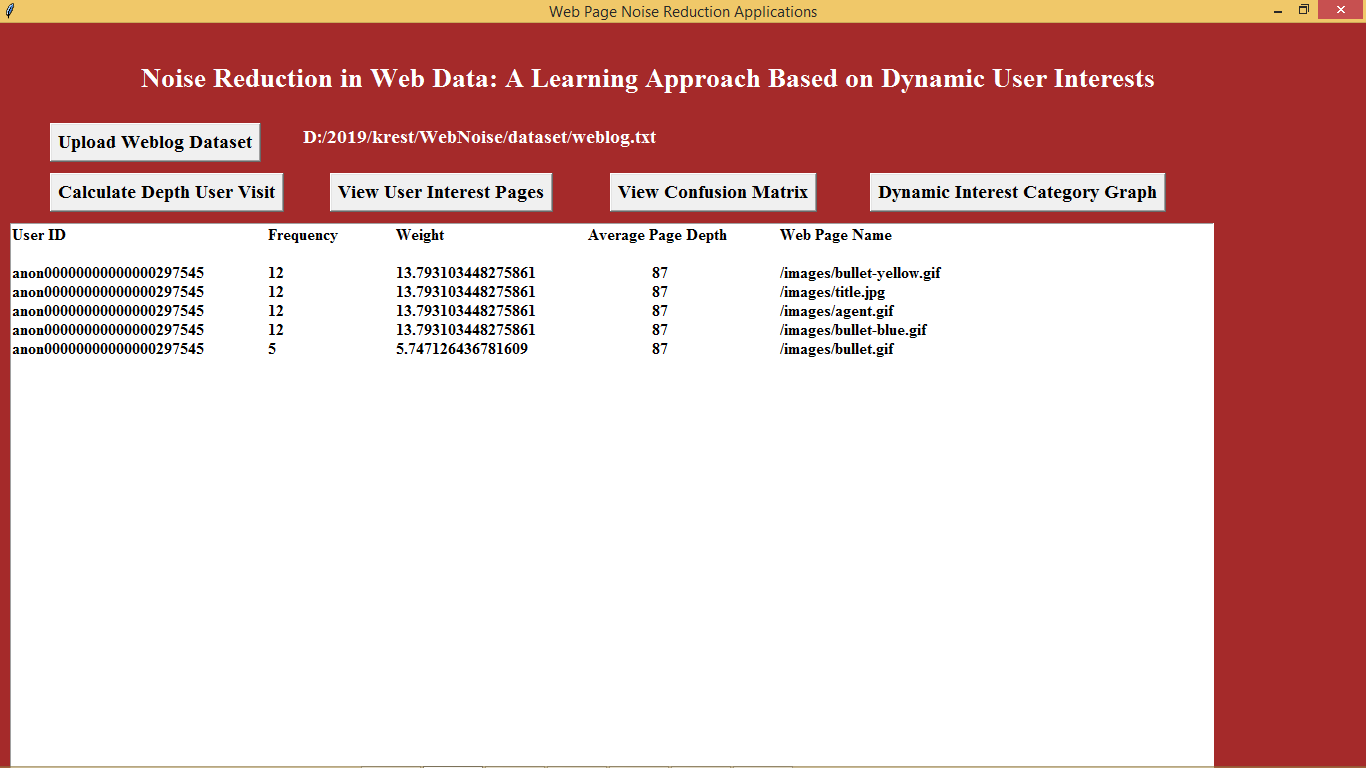


In above screen from command prompt i am copying one user id who access web pages frequently. Now in application click on ‘View User Interest Pages’ button to get below screen

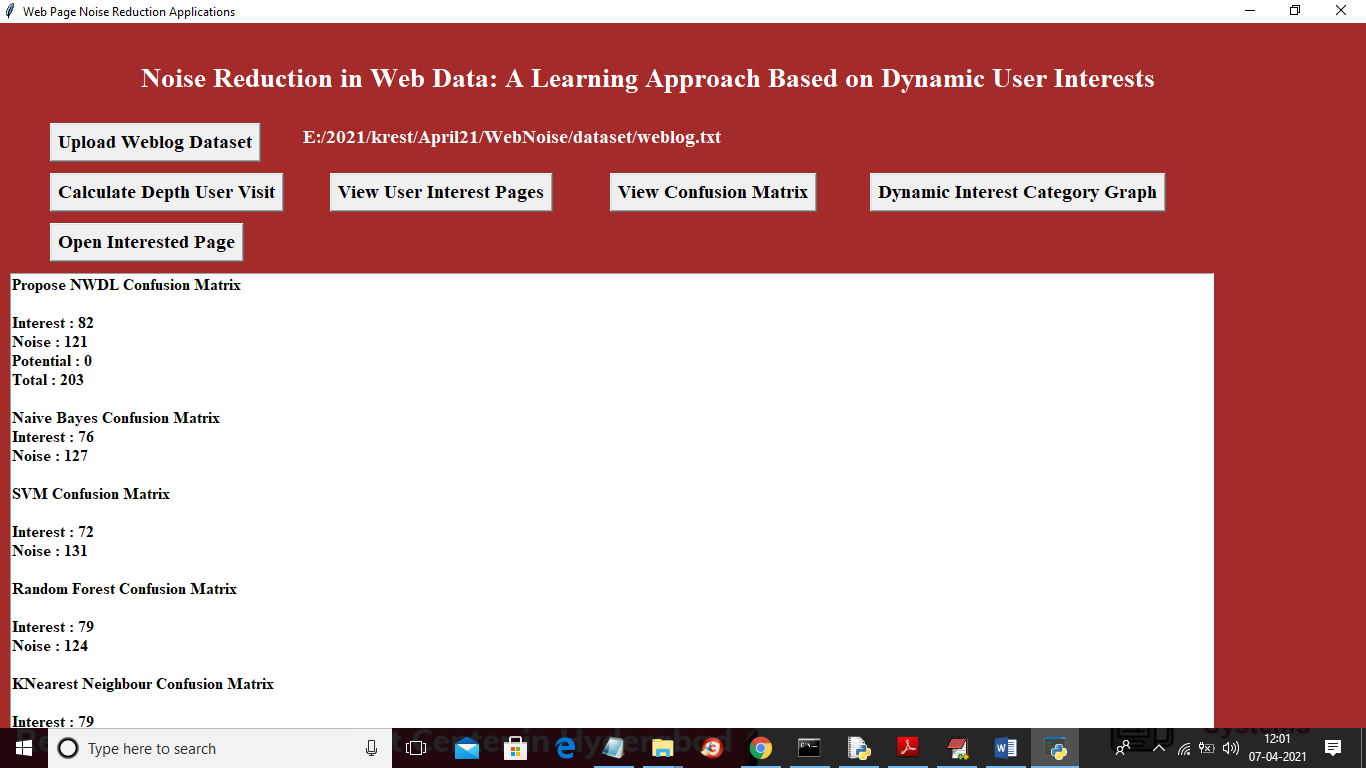


In above screen whatever user id i selected from command prompt i pasted in dialog box. Now click on ok button to get all frequent access pages

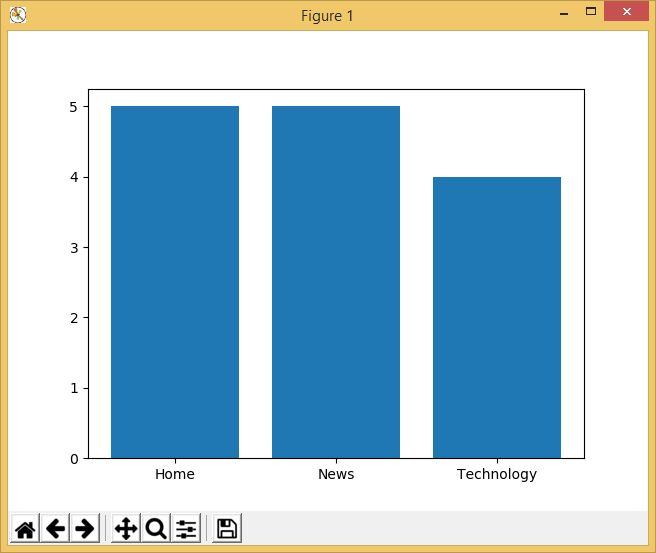




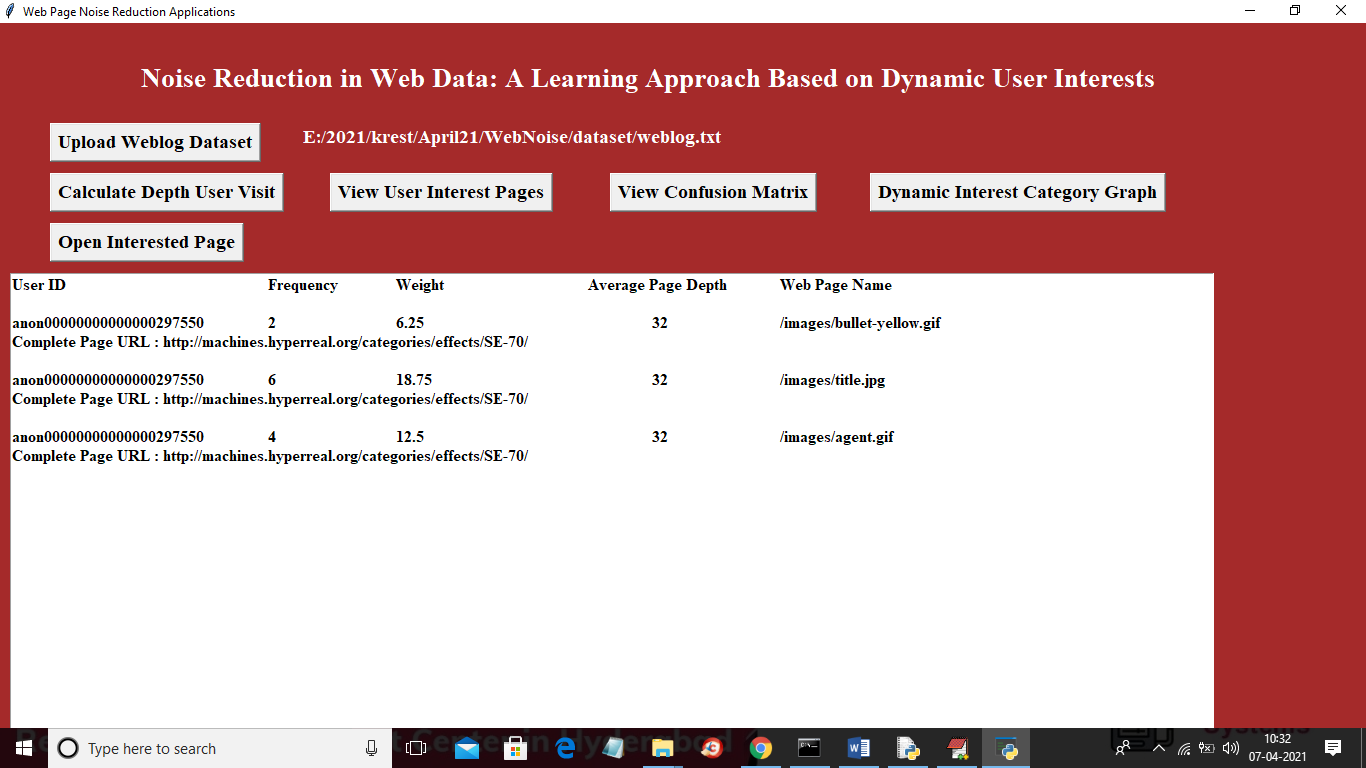
In above screens we can see user interest pages in the ‘Web Page Name’ Column. Above are the web pages details which are access by this user more frequently and will be added to user interested list. Now click on ‘View Confusion Matrix’ button to know no of interested and noise pages obtained by propose NWDL and existing SVM technique



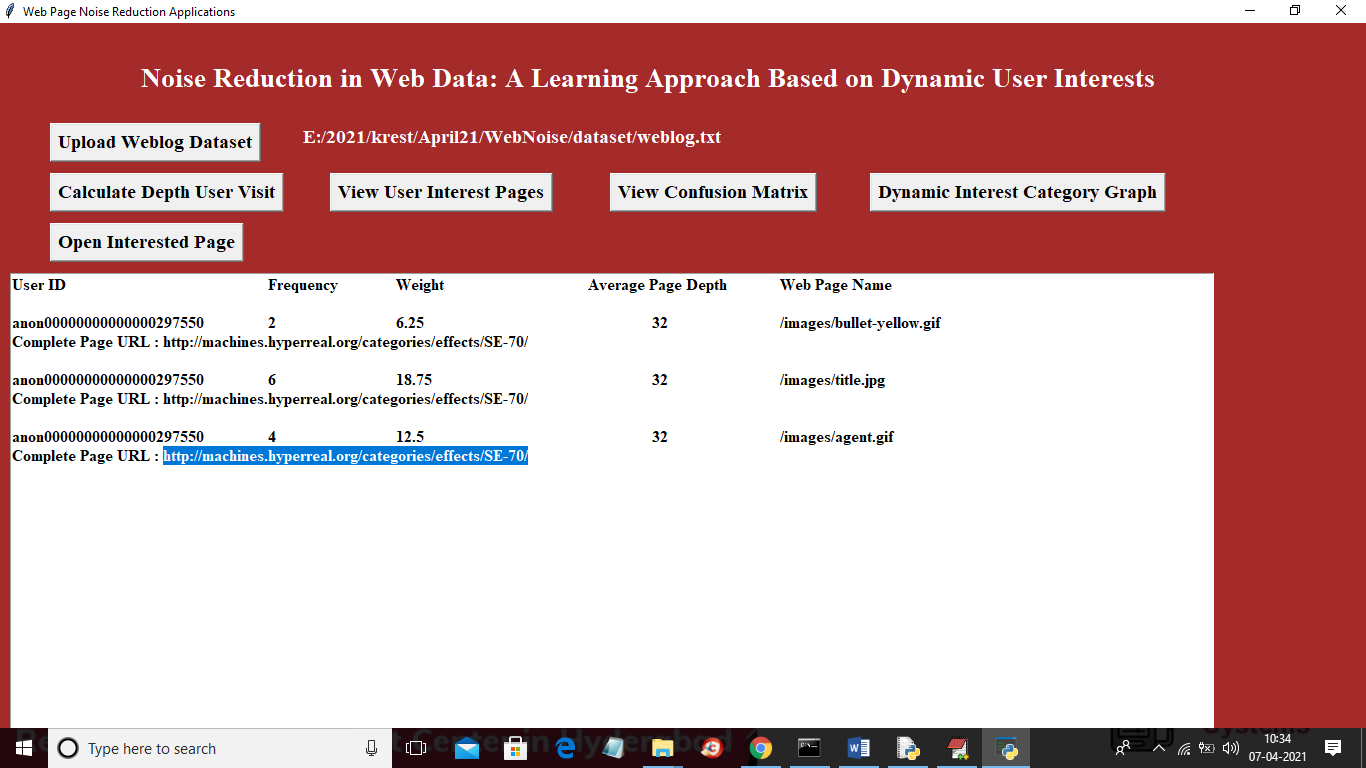
In above screen we can see propose NWDL predict more interested pages compare to existing algorithm like SVM, Naïve Bayes, Random forest and KNN. In above screen with NWDL we got 82% predicted interested pages and naïve bayes we got 72% predicted interested page. Scroll down above text area to view all details of KNN algorithm. Now click on ‘Dynamic Interest Category Graphs’ button to know various categories web pages in the form of graph



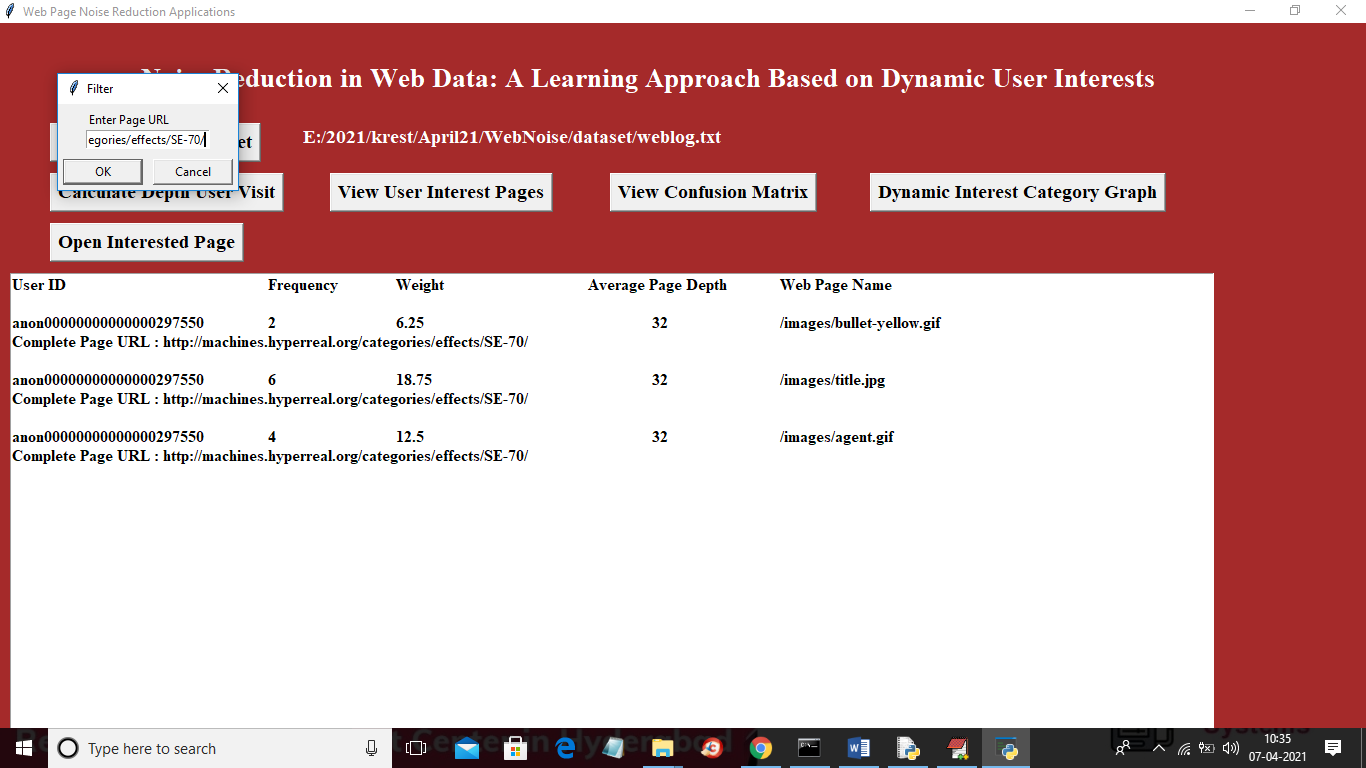
In above screen x-axis represents categories such as Home, News and technology and y-axis represents total count of those categories. Like this in dataset many categories are available but i am showing only 3 categories. Now if user wants to open any interested page then he has to click on ‘View User Interest Pages’ again to get below screen



In above screen user can select any URL like below screen and then click on ‘Open Interested Page’ button



In above screen I selected one URL and then click on ‘Open Interested Page’ button to get below screen



In above screen dialog box paste that URL and click OK button to view that page in browser like below screen

