Analysis on Clickstream Dataset from Wikipedia

Dataset:
I found a dataset from https://figshare.com/articles/Wikipedia_Clickstream/1305770 provided by the Wikipedia. It consists of log information of user navigation from one page ('prev') to another ('curr'). The details of the dataset are given here.

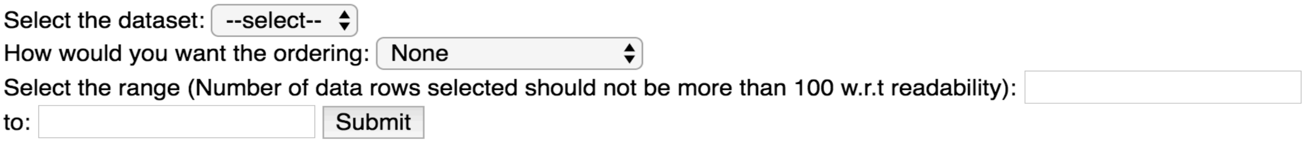
Data description:
Wikipedia Clickstream data is available for January and February of 2015, February, March, April and August of 2016. I will be using February 2016 dataset. Each consists of 4 columns. First 2 columns give information about user navigation from 'prev' page to 'curr' page. Third column tells about the type of navigation:

1. **Link:** wiki-article to wiki-article through links existing in the page.
2. **Other:** wiki-article to wiki-article through a search box i.e., no direct page links between them.
3. **External:** some search engine like google, yahoo or any external website to wiki-article.

The forth column, give the information about number of hits from 'prev' page to 'curr' page. The data for the project is filtered from February 2016 'External' type of pages where users navigated from pages like 'Facebook', 'Google', 'Bing' etc.

Visualization:
Users can pick choices from 2 drop down boxes and 2 text boxes for range in the top of the page.

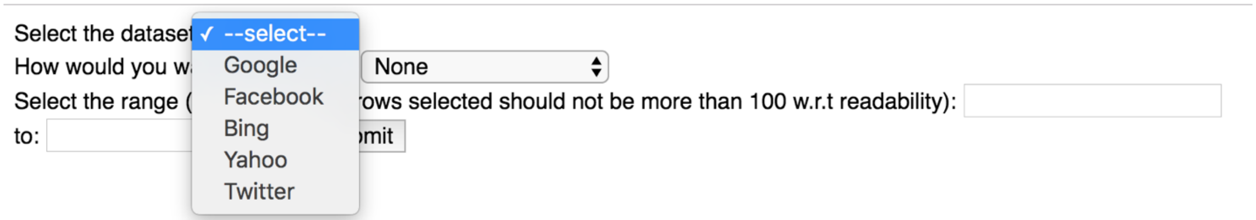
1. Type of 'prev' i.e Google, Facebook, Bing, Yahoo, Twitter
2. Type of sorting order for number of hits.
3. Range of rows from the dataset to be displayed in the visualization.



Select the dataset: --select--
How would you want the ordering: None
Select the range (Number of data rows selected should not be more than 100 w.r.t readability):
to:

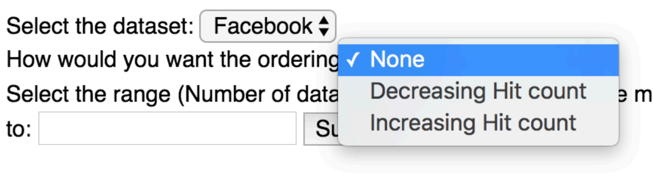
Fig.1. Layout of the choices

Available choices to pick:



Select the dataset
How would you w None
Select the range (ows selected should not be more than 100 w.r.t readability):
to:

Fig .2 Dropdown options for first pick "Select the dataset"



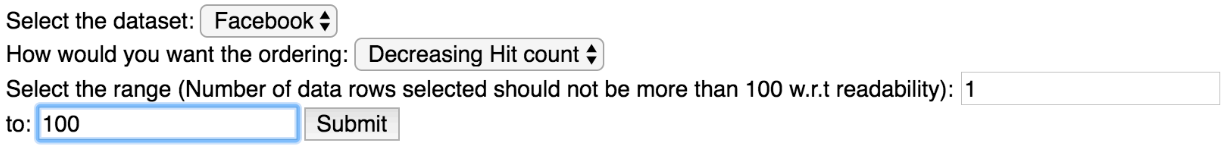
Select the dataset: Facebook

How would you want the ordering

Select the range (Number of data e more than 100 w.r.t readability):

to:

Fig.3 Dropdown to choose any ordering method.



Select the dataset: Facebook

How would you want the ordering: Decreasing Hit count

Select the range (Number of data rows selected should not be more than 100 w.r.t readability): 1

to: 100

Fig. 4 Range of rows to select.

For now, the data range is limited to 100 for better readability of data.
The size of the dataset is displayed below the dropdown selection texts.

The program does not allow the following:

1. Negative numbers are not allowed in the range.
2. Range more than 100 rows is not allowed.
3. First number cannot be greater than second number.

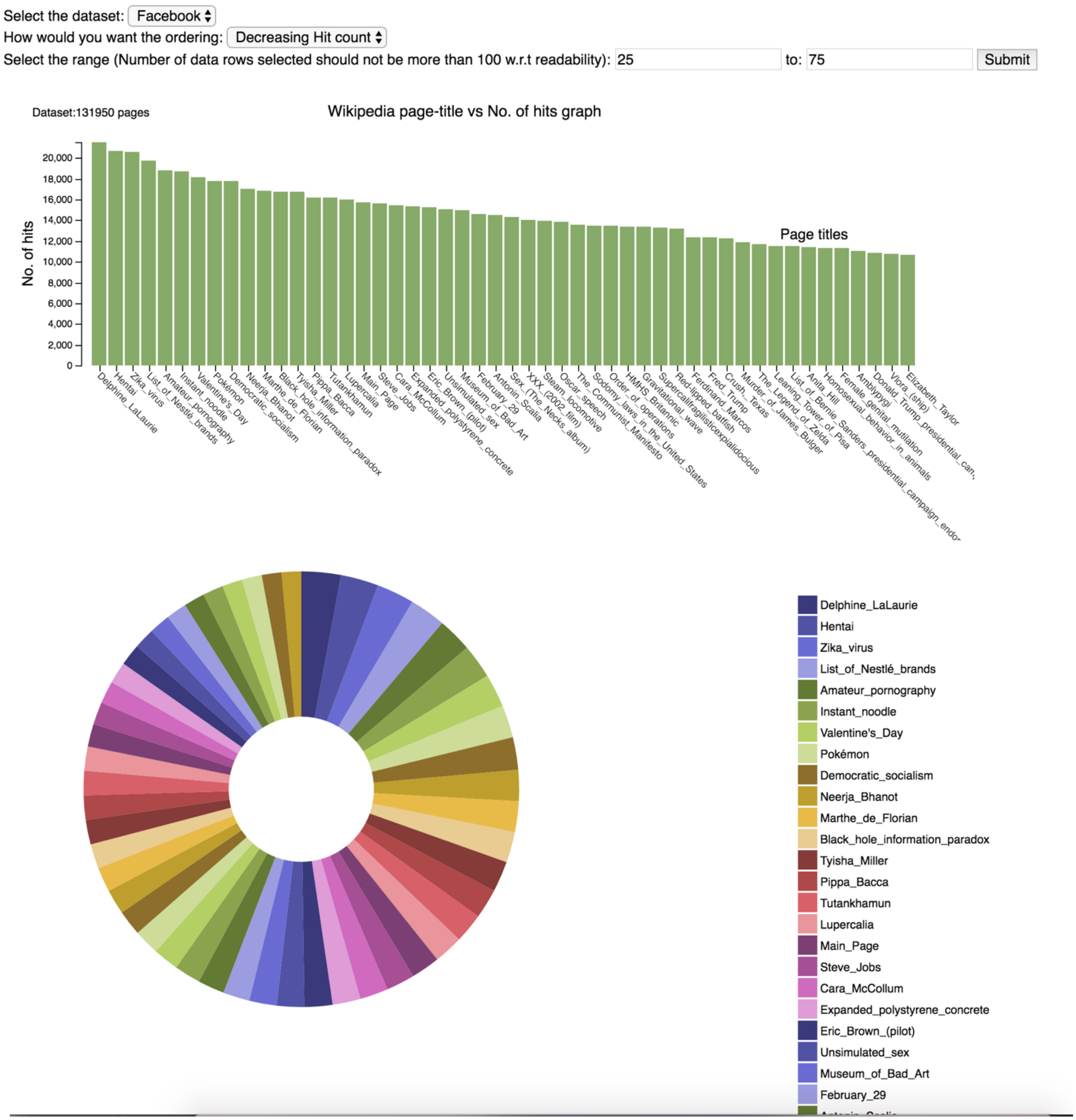
If the first number is left blank, a default 0 is taken. If second number is left blank, a number 100 greater than z is picked. If both are left blank, a random number is picked.

The project visualization consists of 2 d3 layouts.

1. Bar chart: It consists of hit count with respect to each page from the selected dataset. The bars are originally in a visually pleasant green shade. Upon hovering on a bar, it changes to a brown color.
2. Pie chart: The data selected is represented in a pie diagram. Legend on to the right side displays the title in each arc.

Tooltip:

When the cursor is hovered upon a bar or an arc of the pie chart, a tooltip will be displayed with page title, number of hits along with the percentage of count of hits.

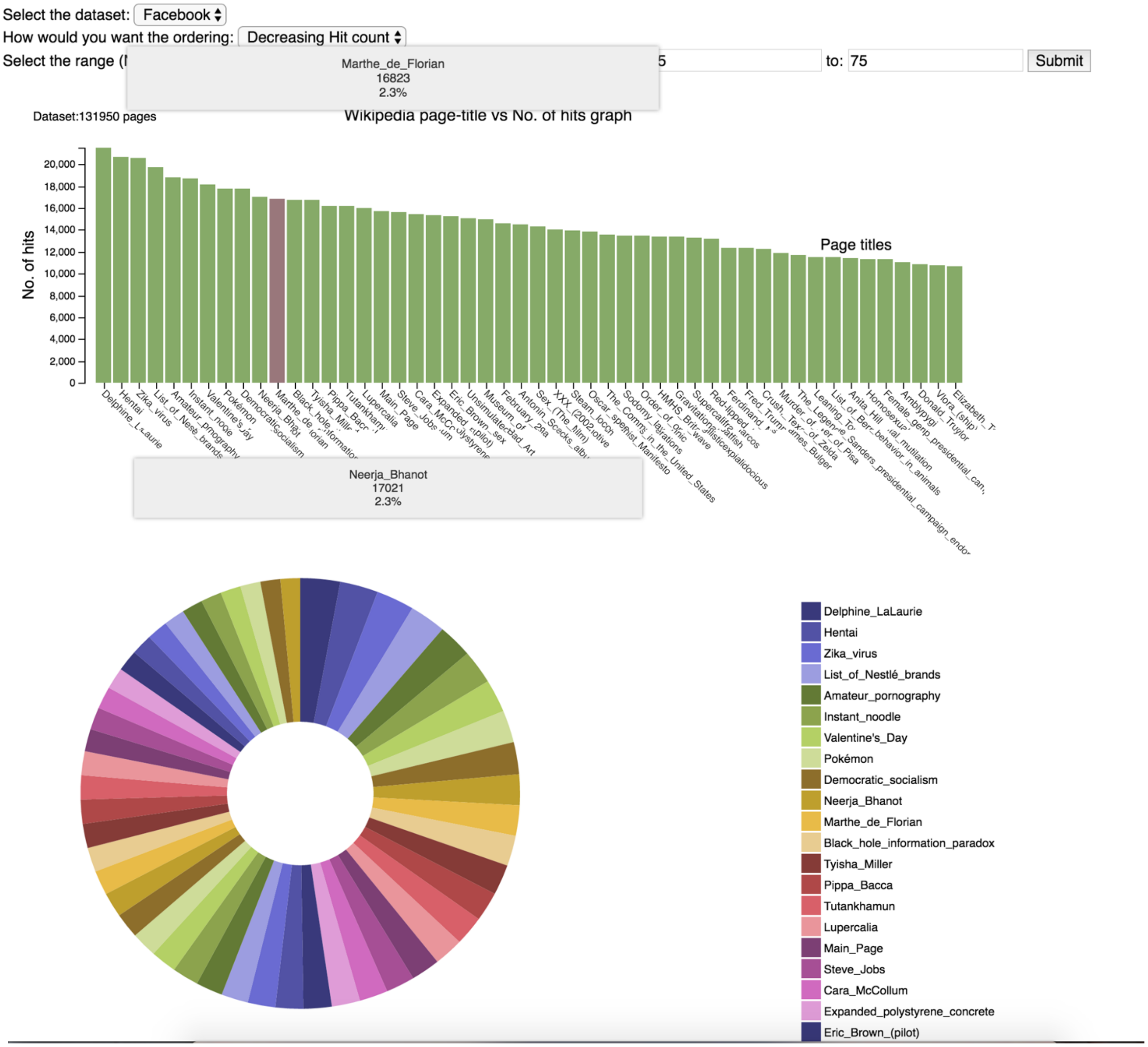


Select the dataset: Facebook ⬍

How would you want the ordering: Decreasing Hit count ⬍

Select the range (Number of data rows selected should not be more than 100 w.r.t readability): 25 to: 75 Submit

Fig: 5 Sample display of the Wikipedia-Clickstream project.



Select the dataset: Facebook

How would you want the ordering: Decreasing Hit count

Select the range (I 5 to: 75 Submit

Marthe_de_Florian
16823
2.3%

Dataset:131950 pages

Neerja_Bhanot
17021
2.3%

Fig. 6. Sample display of the Wikipedia-Clickstream project with tooltip.

Coding:
I used couple of d3 layouts for the visualization. d3 bar chart, d3 Donut pie, d3 legend. I used schemeCategory20b for the pick of colors for pie chart. Based on the choices of drop down, dataset is selected according for loading the diagrams. Sorted according to either increasing hit count or decreasing hit count or neither.

`Data.slice(z,t) –` this can be used to select a chunk of array.

`Data.sort(function(a,b){return a[1] - b[1];}); -` To sort data increasing hit count and vice versa.

`d3.select("#chart_svg").remove(); -` These statements can be used to clear all the svg elements before loading new diagrams. “#chart_svg” is my svg id for pie chart.

Challenges:
There were certain difficulties faced when trying to get the data loaded.

1.Loading big chunks of data
2.Get the toop-tip right.
3.Re-loading the diagrams
4.Placement of legends and sizes.

Reasons for this visualization:

1. To be able to understand the problem of link prediction in Wikipedia and other popular websites.
2. To understand what are the popular Wikipedia pages used by people.

Future Enhancements:

1. More datasets from different ‘types’ in February can be added.
2. More months of 2016 can be included for visualization.
3. A scroll bar can be added for legend, as it is using more space on the web page.
4. Multiple diagrams (bar charts) can be included for different datasets so as to compare hit count of various pages with different ‘prev’ pages like Google, Facebook, Bing etc.