**Cluster Analysis of Web User Sessions**

**Overview:**

* Targeted Advertisement, product/offer campaigns, providing discounts to lure in customer, etc are some of the major application areas of e-commerce.
* All the above mentioned application requires segmentation of customers so as to ease the marketing of products.
* While there are traditionally many ways to group users/customers based on their personal characteristics, this projects aims to categorize them based on their navigational patterns (i.e. pattern in webpages visited by a user) observed from their user sessions.
* Based on the page visits of a user we categorize the users.
* We do a group wise analysis of page hits which can be used to understand the needs of each segment.
* Additionally we build a user network based on sessions and do community detection to identify similar users for targeted advertisements.
* Also a web-page network is built in order identify important pages in a website. Based on the transition probability matrix formed we find the probability of page transitions.

**Dataset:**

**I have provided the dataset in the data folder**

**To learn more about the dataset go to** [**https://archive.ics.uci.edu/ml/datasets/msnbc.com+anonymous+web+data**](https://archive.ics.uci.edu/ml/datasets/msnbc.com+anonymous+web+data)

**I personally had tough time understanding the data**

**So I will boil it down for you guys**

**Go to the above link and under data folder you will find** [msnbc990928.seq.gz](https://archive.ics.uci.edu/ml/machine-learning-databases/msnbc-mld/msnbc990928.seq.gz)

**Download and extract you will get .seq file, you can view it in notepad++**

**It is not much different from .txt, so you can read it normally using “fopen” in python(for reference check the code folder).**

**Dataset explanation:**

**Each line in the dataset represents a user session**

**Assume each line to be a user.**

**Each line is something like this : 2 12 3 4 12 12**

**Meaning user has visited page 2 then 12 followed by 3, 4 and again 12.**

**The website is Microsoft news website**

**It contains pages like sports, medical, economy, accident, education, etc.**

**Each website is denoted by a number**

**So for instance, page 2 means it’s a sports page.**

**Pre-processing of data:**

**Each row represents a user**

**So basically we are gonna cluster each row**

**For that we need a feature vector that represents each user/row**

**So we form a n\*14 matrix**

**Where n is the number of users, 14 is the no of pages in msnbc.com**

**Each user is represented by their page visits.**

**Now we have the dataset ready, deploy all your algorithms using the data.**

**How to check and run the code:**

**Just open the codes under code folder in google colab**

**It has all the required links and documentations.**

**OUTPUT:**

**You can check the final output in index.html page (wait for a couple of seconds to load)**

**Link:**

[**https://nehalmuthu.github.io/-Cluster-Analysis-for-Web-Users-Sessions/**](https://nehalmuthu.github.io/-Cluster-Analysis-for-Web-Users-Sessions/)