**HACKEREARTH – LOVE IN TIME OF SCREENS**

BY RAHUL

Email:rkr.deoghar@gmail.com

18/12/2020

To Execute:

Love\_Screens.py-> This program will implement the below logic

Love\_Screens.ipynb-> This jupyter notebok will help in Analysis of logic

Love\_Screens.html-> Look into to view the logic and implementation

\* NOTE: Since the given data is very large , It will take several hours to produce you a output

Tools Used:

Sklearn- Used for Handling natural language

Pandas- Used for Data Wrangling

Scipy- Used to calculate Euclidean Distances

**APPROACH FOR THE PROBLEM:**

So we have a large data to work on and the problem statement is given with

* User id
* username
* age
* status
* sex
* orientation
* drinks
* drugs
* location
* pets
* smokes
* language
* body profile
* education level
* dropped out
* bio
* interest
* other interest
* location preference

Now let’s try to understand the scenario

We want to find the best matches between the peoples , So the best way to find the people who match each other is to find how similar their preference , likes and dislikes are. So Our approach is:

1. Take two users
2. Compare their characteristics and behaviors
3. Sum up all the compared results
4. Formulate into a equation ***y=mx+c***

**Approach to compare variables:**

**User id**: User id’s only significant is that it is a unique identifier , So we wont be doing any operations on it

**Username**: People don’t judge a person on his name, So we can drop this column

Age: Age is important as people of similar age like to date each other , So for given two persons we find their Euclidean distances as Age\_Euc=Euclidean\_diff(X\_age,Y\_age) and then (1-Age\_Euc/X\_Age) , Note this returns by what factor does age of Y differs from that of X, NOTE : If it is negative we return it as zero

**Orientation and Sex:** These are very much important in deciding the sexual preference of a person

* If a user is heterosexual (prefers the opposite sex), then the match percentage must be **0** for this user with respect to other users of the same gender if the other users have the same behavior.
* If a user is a homosexual (prefers the same sex), then the match percentage must be **0** for this user with respect to other users of the opposite gender if the other users have the same behavior.
* The match percentage of a user with her/himself must be zero.

For more detailed info refer to the code

NOTE: The results with always be a binary i.e 0 or 1

**Drinks , Smokes and Drugs:** People you drink often and those who don’t drink would be a poor match with each other, So first we give each type within this category a Number or we treat this category as ordinal.

We use ***Euclidean Distance*** for calculations

So Two people who Don’t drink at all would have a 100% match i.e 1

Two people One of them who Never drinks and other who often would have a 0% match

Rest them will have a match between 0 and 1 \*exclusive of . Refer to Code for more details

Same goes for Drugs

**Height:** height is of not that much significance also there is no preference regarding height, So that’s ignored

**Job:** If two peoples have same job then it’s a match 1 or else it’s a mismatch 0

**Location and location Preference:** So we have location preference i.e Same State , Same City and Anywhere, for match we get 0 else we get 1

eg- Y ‘s location is LA,California Y’ preference is same state and he lives in San jose , Claifornia then we have a 1

X’s location is LA, California , Y’s preference is same city and he lives in San jose , California then we have a 0 \* For details refer to the code

**Pets:** If likes and dislikes of two peoples match then we have best match i.e 1

We treat this data as Ordinal data and we use Euclidean Distance and then we return the distance.

\* Refer to the code for details

**Language**: In the given data everyone knows English , So we find that if a person wants to learn a new language and the other know a language which the person doesn’t knows , in that case we have a match else it’s a mismatch

**Body Profile**: People prefer partners those are in best shape , so we treat each category as ordinal like Wasted<<Curvy

We return the values \* Refer code for more details

**Education Level:** People have education in range 1-5 , So we return the difference as a percentage in range 0-1.

**Dropped:** People who have a matched values i.e same will have a 1 else 0

**Interests and Other interests:** People have a set of interests , So find a set intersection of them which are common, and we return the percentage of ,matched in range 0-1

**Bio:** Bio uses a much complex NLP algorithms , We take Bio of two people and we count vectorize the words and then we find the cosine similarity. \* Refer to the code for details

**Calculating Data:**

**Iterate over the users and compare them to others- Complextiy of Iteration O N^2**

Now All calculated Now Refer to the Sexual Orientation section-

So we can say that Sexual Orientation decides wheat to Percentage of match is 0 or 1 , So it is slope rather than a intercept or variable. Like that of location preference .

So we have equation as :

(O1\*O2)\*(O2+O3+O4+O5+O6+O7+O8+O9+O10+O11+O12+O13+O14)\*(100/12)+0=Y

Mx+C=y

**NORMALIZATION \* NOT MANDATORY**

Here we have C as zero

No we will have a score in range 100-0 inclusive of

So now we need to Normalize the data , so that the best Matched will have a value 100% or to uniformly spread that data in range 0-100

NOTE: Sometimes , because the Data is large and takes a lot of memory , Some values may come as negative , So handle those .

Write the Data to CSV file in a proper format