ROOM8

SEGMENTATION OF ROOMMATES ON SOCIAL MEDIA PLATFORMS

SOWMIYA MURUGANANDAM

BUSINESS Problem

Identify market segments on social media and find out potential roommate match

Our Approach

Data Mining and attribute extraction from Twitter

Cluster Market segments using Evolutionary Solver

Match users using Linear Optimization

Twitter Data Mining

Mined Potential users looking for roommates on Twitter

Used a Hashtag based search

Used NLP to classify users into attributes

Categorized users into five attributes

- Clean
- Night Owl
- Student
- Smoker
- Pet Owner

Clustering using Evolutionary solver

Inputs:

i: index representing users, where $i \in \{1,2,..50\}$

j: index representing cluster, j $\in \{1,2,..K\}$

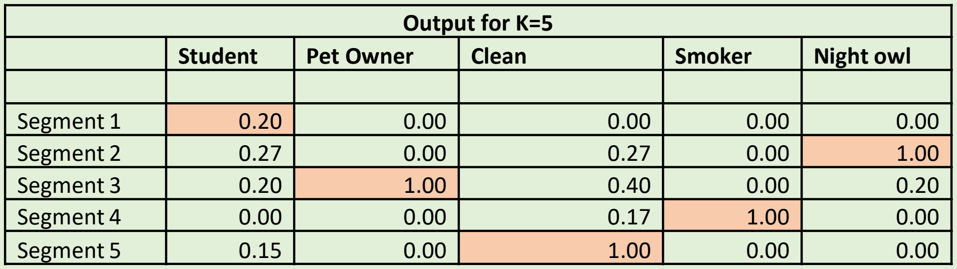
c_j: cluster center

x : user vector

K: number of cluster

Decision Variable:

c_j: cluster center



Output for K=5					
	Student	Pet Owner	Clean	Smoker	Night owl
Segment 1	0.20	0.00	0.00	0.00	0.00
Segment 2	0.27	0.00	0.27	0.00	1.00
Segment 3	0.20	1.00	0.40	0.00	0.20
Segment 4	0.00	0.00	0.17	1.00	0.00
Segment 5	0.15	0.00	1.00	0.00	0.00

Constraints: $(1)\ c_j \in integer$ (2) $c_j \leq 50$ (3) $c_j \geq 1$

Objective: To minimize the squared error

$$\mathbf{min}(\textstyle\sum_{\boldsymbol{i=1to\ 50}} \quad \sum_{\boldsymbol{j=1\ to\ k}} (||\ \boldsymbol{xi} - \boldsymbol{cj}\ ||)\mathbf{2})$$

Stable Roommate Match

Inputs:

i: index representing users in set 1, where $i \in \{1,2,..50\}$

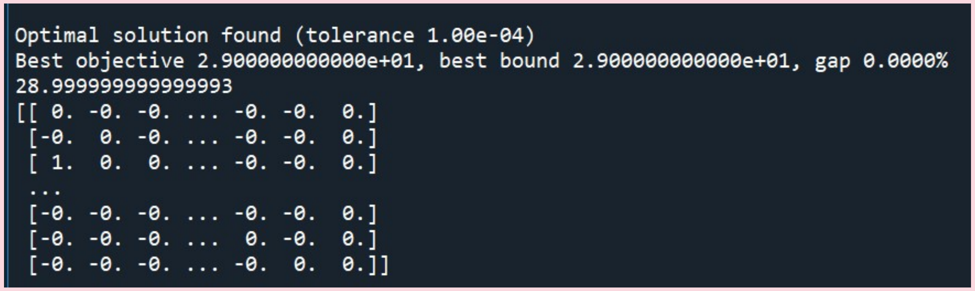
j: index representing users in set 2, j $\in \{1,2,..50\}$

d_{ij}: hamming distance between the users

x_i, x_j: Users

Objective: minimize distance between users

$$\mathbf{min(} \sum_{i=1 to\ 50} \sum_{j=1\ to\ 50} (\boldsymbol{a_{ij} * d_{ij}})$$



```
Optimal solution found (tolerance 1.00e-04)
Best objective 2.900000000000e+01, best bound 2.900000000000e+01, gap 0.0000%
28.999999999999993
[[ 0. -0. -0. ... -0. -0.  0.]
 [-0.  0. -0. ... -0. -0.  0.]
 [ 1.  0.  0. ... -0. -0.  0.]
 ...
 [-0. -0. -0. ... -0. -0.  0.]
 [-0. -0. -0. ...  0. -0.  0.]
 [-0. -0. -0. ... -0.  0.  0.]]
```

Decision Variable:

a_{ij}: binary decision variable

Constraints:

(1) $a_{ij} \in \{0,1\}$ (2) $\sum_{j=1}^{50} a_j = 1$ (3) $\sum_{i=1}^{50} a_i = 1$

Caveats

For the clustering model, there are certain limitations as below,

1) Conventional solvers cannot handle the increase in user size, and processing time might be an issue going forward
2) To find the optimal K value, we had to run a number iterations, which could be eliminated if we included optimal calculation of K Value as part of the model
3) Automatically fetching and clustering large scale data on a time-on-time basis will be a problem with the existing model
4) We could also try other approaches like convex clustering to compare the best results

For the Matching model,

1) Increase in user base will increase the processing time.

Thank You!