Phase 2 – Team 3 – Mid-Project Report

[The initial results after completing preliminary analysis]

Members: Hanyu Chen, Yipeng Guo, Luke Hong, Ting-Ann Lu, Yuyan Ma

A) - The way we do **data cleaning** and how we propose to apply unsupervised machine learning methodology based on it.

- Set an ID for each user and use it as an index, this will help us with matching.
- Drop Columns: last online/ speaks/ jobs/ sign, which do not contribute to our recognition of characteristics.
- Replace reasonable NaH values to reduce the null value proportion.
- Drop the rest of entries containing null values.
- Combine all the essays into one as 'Full_essay'; Keep location & religions as the three text components we are gonna analyze.
- Rest of them we either make them into dummy or give them a scale.

After this kind of manipulation, we have a dataframe shaped (9055, 20). Separate the dataset into two: **Text** one contains 'Full_essay' 'location' and 'religions' ; **Numerical** one contains the rest of the columns.

B) - Basic findings after implementing multiple machine learning tools.

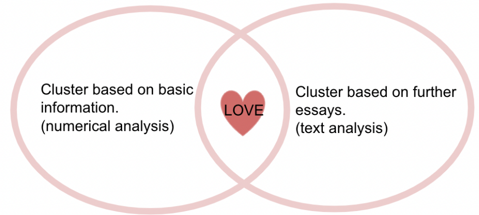
- **PCA**
    - PCA for numerical variables
    - Variables have low correlation
        - Principal components did not contribute to reducing noise
        - Age and offspring have highest correlation if we want to consolidate in terms of dimensionality but does not affect our results by much
- **H-Clustering**
    - Based on PCA, 11 variables are used in the Hierarchical Clustering model.
    - We choose our linkage method of "ward" since it is more balanced and faster.
    - Based on the dendrogram, we select 5 and 8 groups for our clustering numbers.
        - 5 groups - Average silhouette score of 0.195; The distribution is uneven and cluster 3 has many negative values.
        - 8 groups - Average silhouette score of 0.163; The distribution is getting better, but cluster 4, 6 and 8 still have many negative values.
    - Based on the silhouette score plot, we believe that the best number of H-clustering should be **5**.

- **K-Means Clustering**
    - Based on PCA, 11 variables are used in the K-means clustering model
    - Inertia plot and Silhouette score plot are used to make the decision of the K
        - The inertia with number of clusters plot illustrates the elbow pattern where the optimal solution for k lays between 9 to 15
        - The peak point of the silhouette plot lies on the k of 18.
        - **K of 16 is chosen after trading off between the two plots.**
    - Fit our numerical data with the K-means model
        - By looking at the cluster distribution with a silhouette score plot. We get the average silhouette score of 0.227.

- **Text Analysis - Basic CV & Sentimental Analysis**
    - CountVectorizer
        - Without any tuning got 128806 tokens as result
        - Removed stopwords reduce tokens to 128676
        - Then we transform the dataset into data frame the shape is (9055, 128676
        - Colab can not generating visualization so we tune the model and apply PCA (Setting max features = 10000, PCA (components = 100)
            - We can explain 82% of variance with 100 components
    - Combined TweetTokenizer with CountVectorize
        - The combination result in a more efficient way to handling data
        - It was the only model that colab can visualize the result.
    - Applied Tfidfvectorizer (removed stopwords) and receive the same shape of data frame as CV
    - Applied sentimental analysis and find out our users have their specific characteristics - positive/negative, introvert/outgoing etc.

C) - **Next Step:** What are we going to do to realize our matching system?

Based on our former analysis, we make it to cluster our users based on their basic information (specifically speaking, the numerical columns), hence we have cluster1. We can also fit a prediction model to cluster newbie users.


Cluster based on basic information. (numerical analysis)
LOVE
Cluster based on further essays. (text analysis)

We also want to do another cluster based on text analysis, to find out some outstanding characteristics. Once the newbie users input their basic info, they will be distributed into the basic user pool. Then they will be shown a question about 'What characteristic do you crave most?' together with those ones we extracted. Just click on it and our system will guide you to find the ones with the specialty you like, and also in the similar user pool with you. Here possibly comes love! ❤️