Big 5 personality traits

Alexia

Data Exploration

```
> summary(data)
   X.AUTHID            STATUS               X                X.1              X.2              X.3
 Length:9916        Length:9916        Mode:logical     Mode:logical     Mode:logical     Mode:logical
 Class :character   Class :character   NA's:9916        NA's:9916        NA's:9916        NA's:9916
 Mode  :character   Mode  :character

   X.4               DATE            NETWORKSIZE      BETWEENNESS        NBETWEENNESS       DENSITY
 Mode:logical     Length:9916        Min.   :  24.0   Min.   :     185.7  Min.   :31.21   Min.   :0.00000
 NA's:9916        Class :character   1st Qu.: 196.0   1st Qu.:   16902.2  1st Qu.:93.77   1st Qu.:0.01000
                  Mode  :character   Median : 317.0   Median :   47166.9  Median :96.44   Median :0.02000
                                     Mean   : 426.4   Mean   :  135439.0  Mean   :94.67   Mean   :0.03029
                                     3rd Qu.: 633.0   3rd Qu.:  196606.0  3rd Qu.:97.88   3rd Qu.:0.03000
                                     Max.   :1596.0   Max.   :1251780.0   Max.   :99.82   Max.   :0.40000
   BROKERAGE        NBROKERAGE     TRANSITIVITY         cEXT                cNEU
 Min.   :    241   Min.   :0.32   Min.   :0.0000   Length:9916        Length:9916
 1st Qu.:  17982   1st Qu.:0.49   1st Qu.:0.0600   Class :character   Class :character
 Median :  48683   Median :0.49   Median :0.0900   Mode  :character   Mode  :character
 Mean   : 137656   Mean   :0.49   Mean   :0.1288
 3rd Qu.: 198186   3rd Qu.:0.50   3rd Qu.:0.1700
 Max.   :1263790   Max.   :0.50   Max.   :0.6300
     cAGR               cCON               cOPN
 Length:9916        Length:9916        Length:9916
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

STATUS
likes the sound of thunder.
is so sleepy it's not even funny that's she can't get to …
is sore and wants the knot of muscles at the base of …
likes how the day sounds in this new song.
is home. <3
www.thejokerblogs.com

Task: to cluster the 250 users and explore the relationship between Big 5 personality and each cluster

Data Wrangling

STATUS
likes the sound of thunder.
is so sleepy it's not even funny that's she can't get to …
is sore and wants the knot of muscles at the base of …
likes how the day sounds in this new song.
is home. <3
www.thejokerblogs.com

- num_characters: The number of characters in each post.
- num_punc: The number of punctuation marks used in each post (including !, ~, and #).

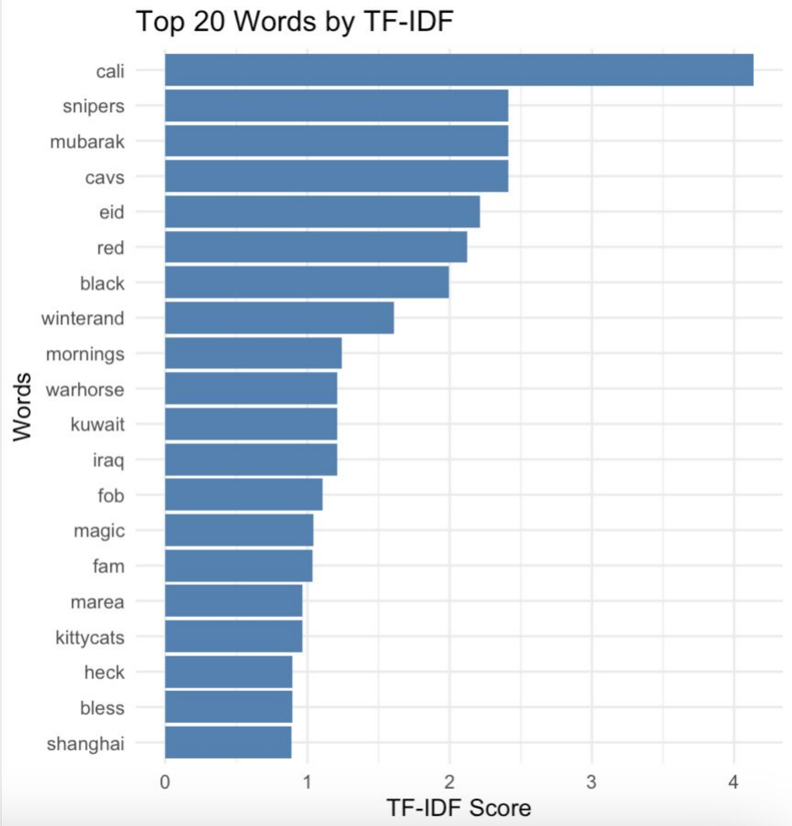
Data Wrangling

STATUS
likes the sound of thunder.
is so sleepy it's not even funny that's she can't get to …
is sore and wants the knot of muscles at the base of …
likes how the day sounds in this new song.
is home. <3
www.thejokerblogs.com

lowercase,
remove punc,
remove stop words,
tokenise

word
likes
sound
thunder
sleepy
funny
sleep
sore

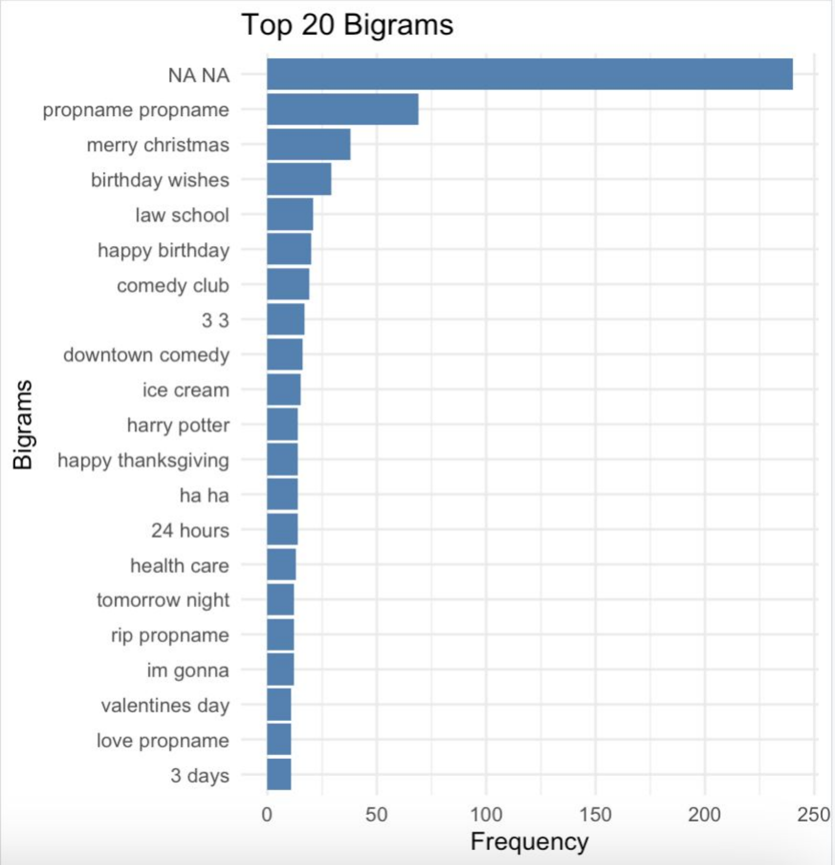
Data Wrangling– TF-IDF



Term Frequency Inverse Document Frequency:

How important a word is to a document in a collection (or corpus) of documents

Data Wrangling – N-grams



Data Wrangling – Sentiment Analysis

TheMysMan_bh@d4jk:

“My friend Kanye **punched** me because I cheated on him with his girlfriend Kim, but I still feel **happy** today for Kim’s **kiss** ! xoxo!!!”

Positive ratio: 3/(1+3)

Data Wrangling

```
$ cOPN           : chr  "y" "y" "y"
$ num_punc       : int  0 0 1 0 0 0
$ num_characters : int  26 59 116 4
$ mean_tf_idf    : num  0.00634 0.0
$ top_word       : chr  "3" "3" "3"
$ positive_ratio : num  0.637 0.637
```

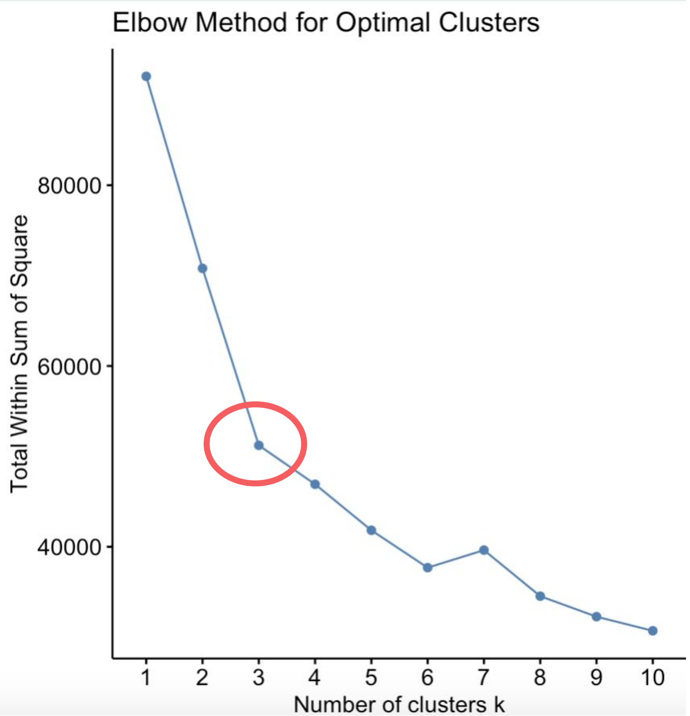
Data Wrangling

🌲🌲🌲

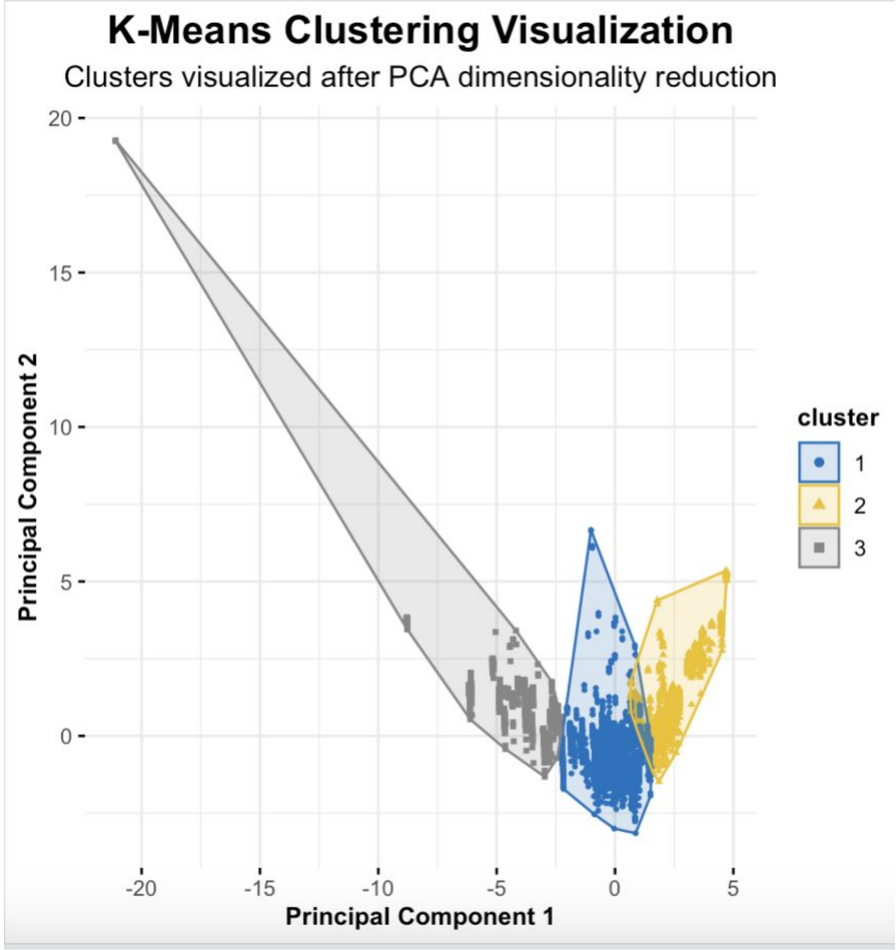
RF for feature importance (eg. **cEXT**)

	n	y	MeanDecreaseAccuracy	MeanDecreaseGini
NETWORKSIZE	11.023787	10.437391	12.831892	185.0561380
~~BETWEENNESS~~	~~11.447184~~	~~11.205105~~	~~13.358052~~	~~185.7450730~~
NBETWEENNESS	12.154502	10.844694	13.555254	171.8532026
NBROKERAGE	6.382967	7.028083	8.207925	70.0791159
DENSITY	8.588331	5.830041	9.444128	45.7282323
TRANSITIVITY	12.670610	11.238558	13.815618	249.3083881
num_punc	2.837492	1.181884	2.923152	0.4568379
num_characters	2.248371	1.396597	2.541610	0.6145732
mean_tf_idf	14.767135	10.831288	14.233036	159.0392216
positive_ratio	11.631051	11.407710	12.691495	88.4420601

K-Means

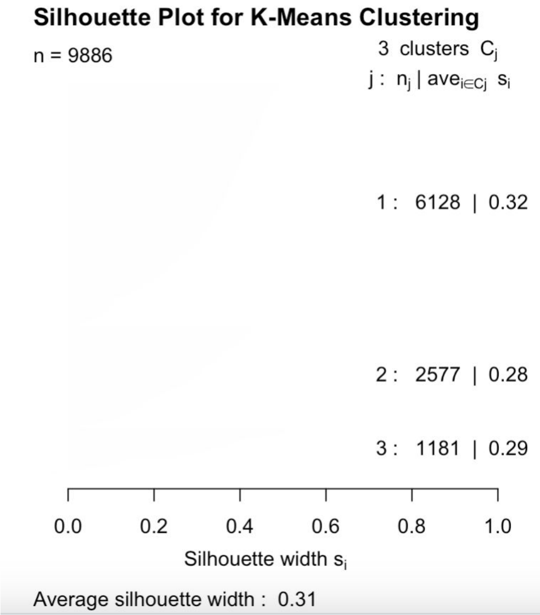


Elbow Method for Optimal Clusters
Total Within Sum of Square
80000
60000
40000
1
2
3
4
5
6
7
8
9
10
Number of clusters k



K-Means Clustering Visualization
Clusters visualized after PCA dimensionality reduction
Principal Component 2
20
15
10
5
0
-20
-15
-10
-5
0
5
Principal Component 1
cluster
1
2
3

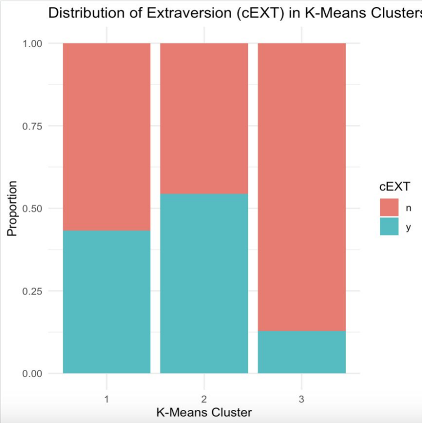
K-means – evaluation

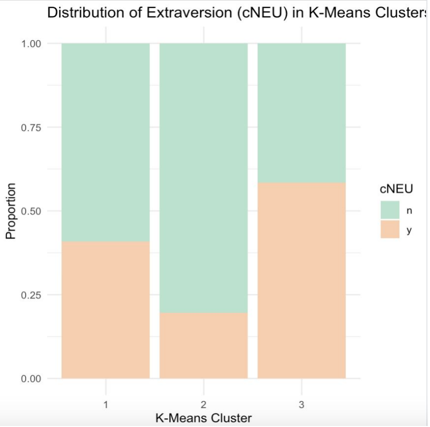


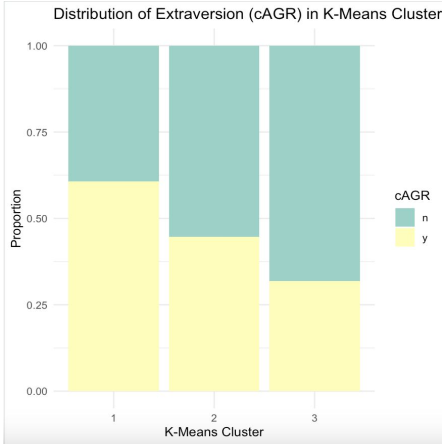

```
> purity_cEXT
> print(purity
[1] 0.6613892
> purity_cNEU
> print(purity
[1] 0.6598889
> purity_cAGR
> print(purity
[1] 0.6143171
> purity_cCON
> print(purity
[1] 0.59881
> purity_cOPN
> print(purity
[1] 0.74366
```

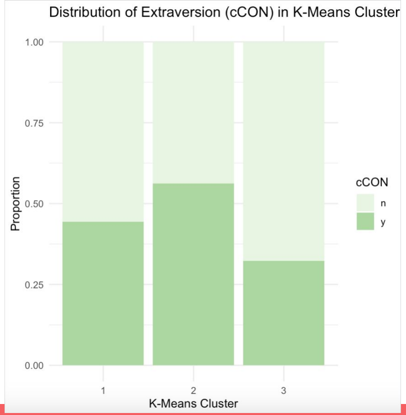
a measure of cluster quality in relation to each personality

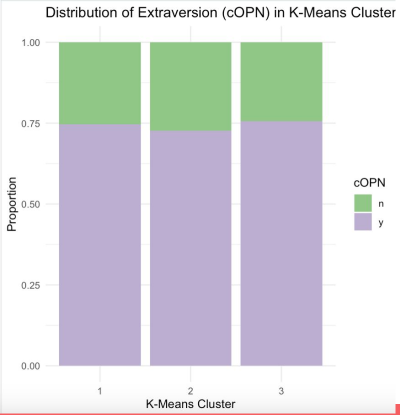
K-means evaluation



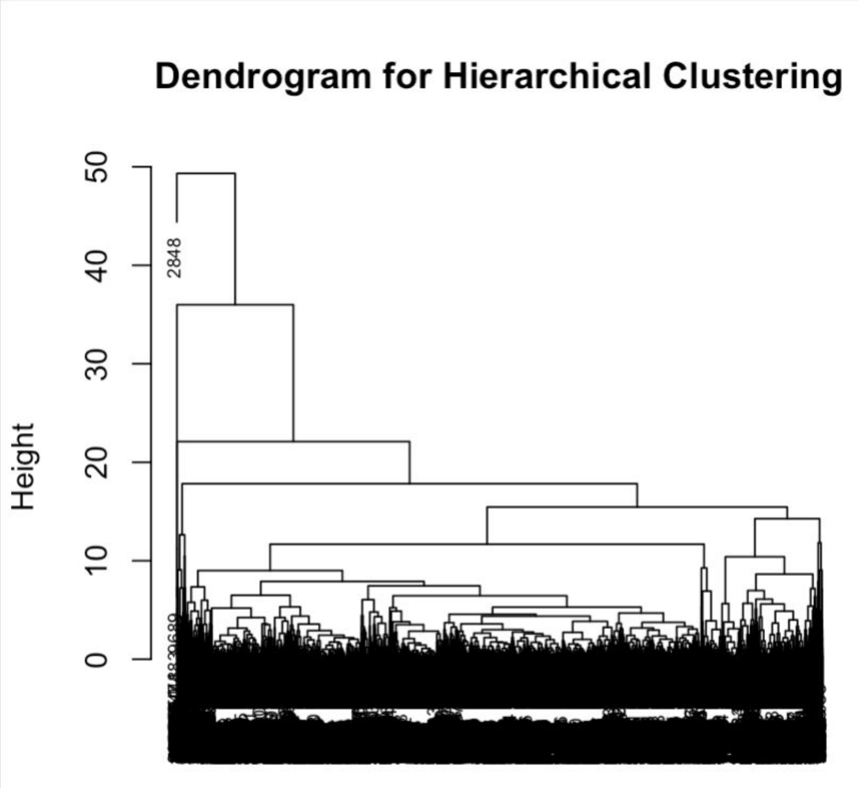








Hierarchical Clustering

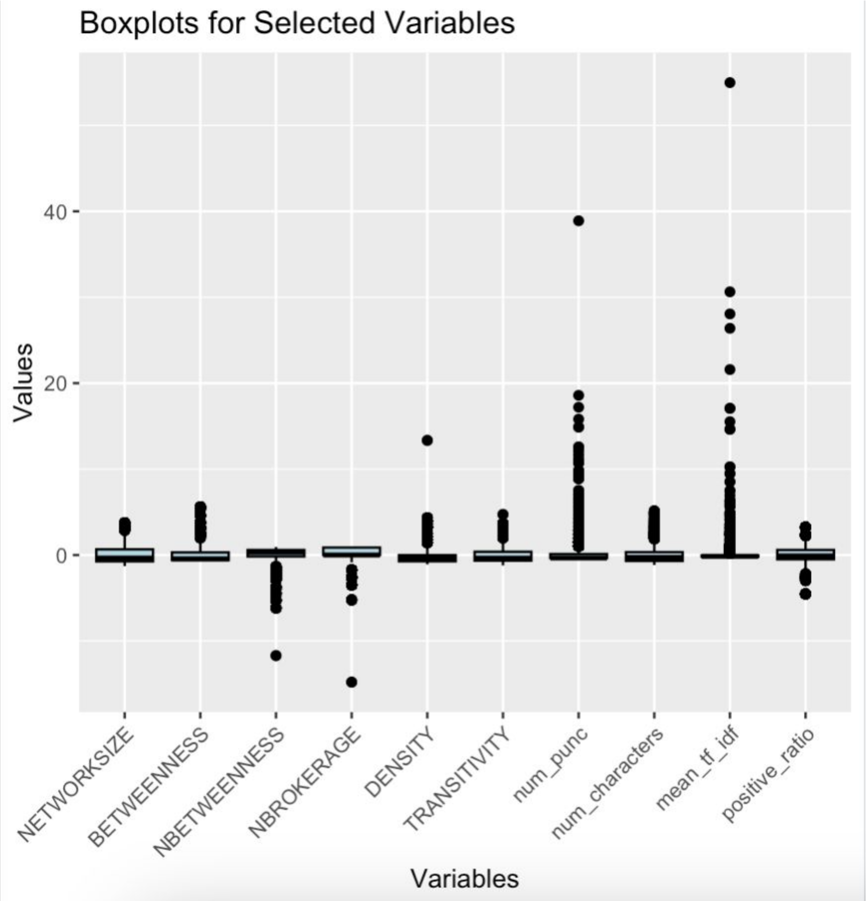


Hierarchical Clustering — evaluation

```
> print(Hmean_silhouette)
[1] 0.8726472
```

Well-separated pt;
Compact clusters —- minimal overlap between

```
> print(paste("Purity for cEXT:", purity_
[1] "Purity for cEXT: 0.858140240817566"
> purity_cNEU <- purity_score(dw_clean$h
> print(paste("Purity for cEXT:", purity_
[1] "Purity for cEXT: 0.875341495497319"
> purity_cAGR <- purity_score(dw_clean$h
> print(paste("Purity for cAGR:", purity_
[1] "Purity for cAGR: 0.843940773719181"
> purity_cCON <- purity_score(dw_clean$h
> print(paste("Purity for cCON:", purity_
[1] "Purity for cCON: 0.846639009747378"
> purity_cOPN <- purity_score(dw_clean$h
> print(paste("Purity for cOPN:", purity_
[1] "Purity for cOPN: 0.914938109211103"
```


Boxplots for Selected Variables
Values
40
20
0
Variables
NETWORKSIZE
BETWEENNESS
NBETWEENNESS
NBROKERAGE
DENSITY
TRANSITIVITY
num_punc
num_characters
mean_tf_idf
positive_ratio

Advantages & disadvantages

– Advantages

1. Integration of Structured and Unstructured Data
2. Model Evaluation
3. Real-World Application

– Disadvantages

1. Text Complexity
2. Moderate K-Means Performance
3. Feature Selection Limitations

Potential Challenges and Considerations

1. Data Overlap

   overlapping nature of personality traits
   (e.g., Agreeableness and Conscientiousness)

2. Scalability

3. Text Data Variability

   Variations in language usage, slang, and punctuation

Ethical considerations

1. User Privacy

   Maintaining anonymity and protecting users from re-identification is crucial

2. Data Sensitivity

   Social media data can be misused in harmful ways

3. Algorithmic Fairness

   Ensure that clustering methods do not reinforce biases or stereotypes

THANK

YOU !