Parker Williamson
3/21/2018
Springboard Data Science Career Track

Capstone Project 2 Milestone 1 – Text Mash

Problem

Improve interactions between people and improve user experience. Matching people and finding fun activities are very labor intensive manual process and very important for people's life experiences. Understand who people are and what they are likely to want is one of the most important things to identify to really make a difference in most people's lives.

Client

Any apps would be able to use the personality traits to improve the experience. Personality traits can be used to pair like-minded people for dating, assistance or finding what activities all people in a group would be most likely to enjoy based on all of their personalities. It could also be used to present information in the way each type of person finds most appealing. It could even be used to understand how an email would make you seem (optimistic, extroverted, ext). If the solution is strong enough it could be useful enough to start a business around.

Dataset

I will focus on accurately predicting MBTI personality types and then expand to other personality features such as positivity/negativity if time allows. The MBTI raw data I will get from Kaggles Myers Briggs dataset (https://www.kaggle.com/datasnaek/mbti-type). Some personality types don't post as much so I will try to find more data to make a large enough sample, so that can have more examples of the types less vocal on forums. That extra data I will collect from Reddit's MBTI subbreddits which have posters self labelled type and separate subreddits for different types. The subreddits for different types could be used if I filter out any comments that include references to other personality types.

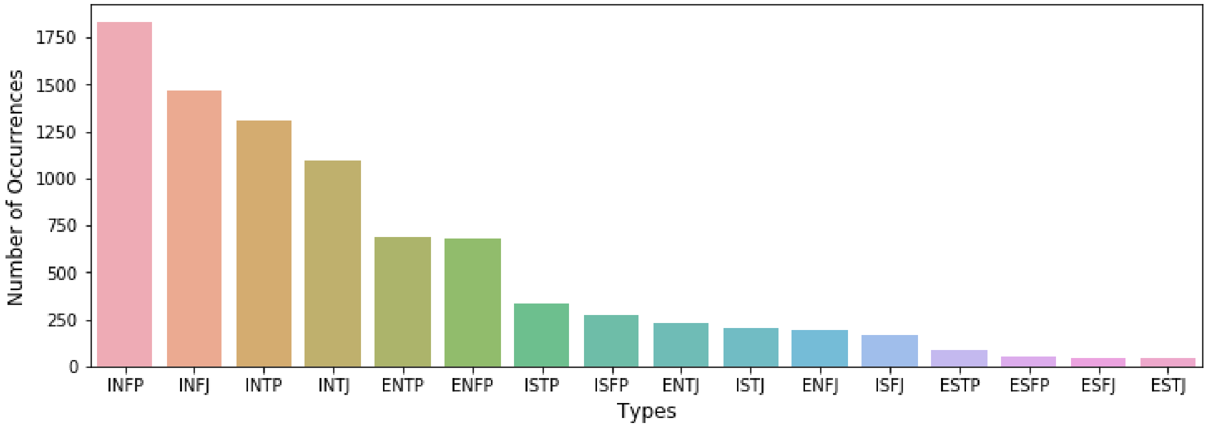
Alternative datasets

If time allows I could add a number of other features to help get a full personality or linguistic profile:

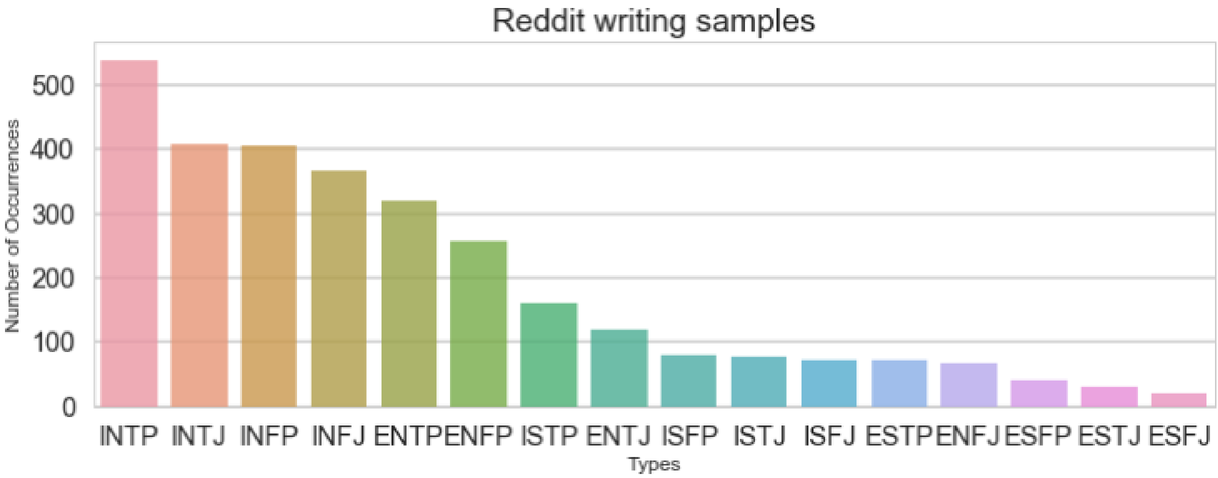
Emotion (https://www.kaggle.com/c/sa-emotions/data)
Topic of conversation (https://www.kaggle.com/c/comp-551-miniproject-2-reddit-classification)

Findings

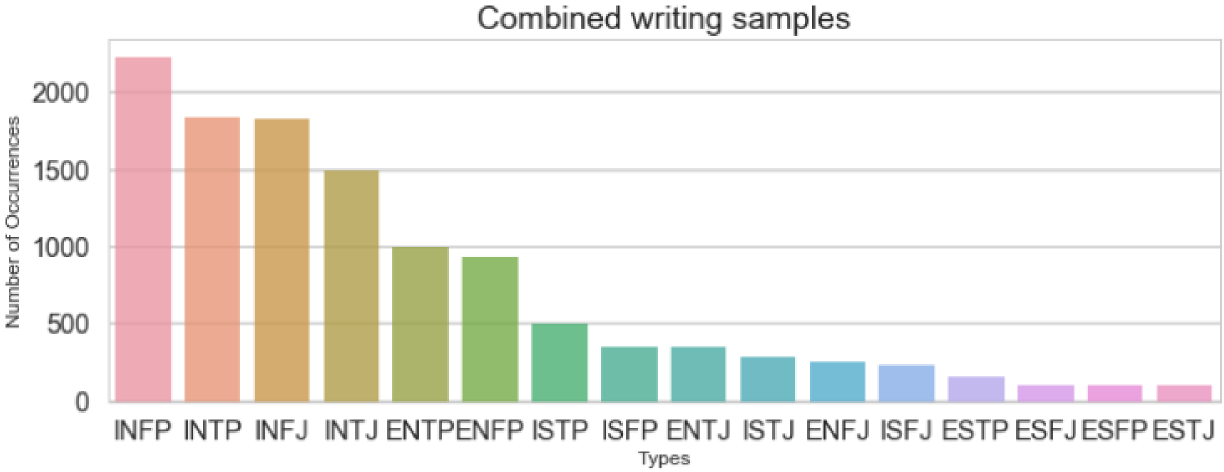
The initial Kaggle dataset is biased towards introversion, and so are the subreddits which means some personality types are more likely to post on forums. Although that means that the personality traits are more likely to influence what people enjoy, that makes it harder to collect an unbiased dataset. To try and achieve an unbiased dataset I combined the Kaggle dataset with the data I collected from reddit.



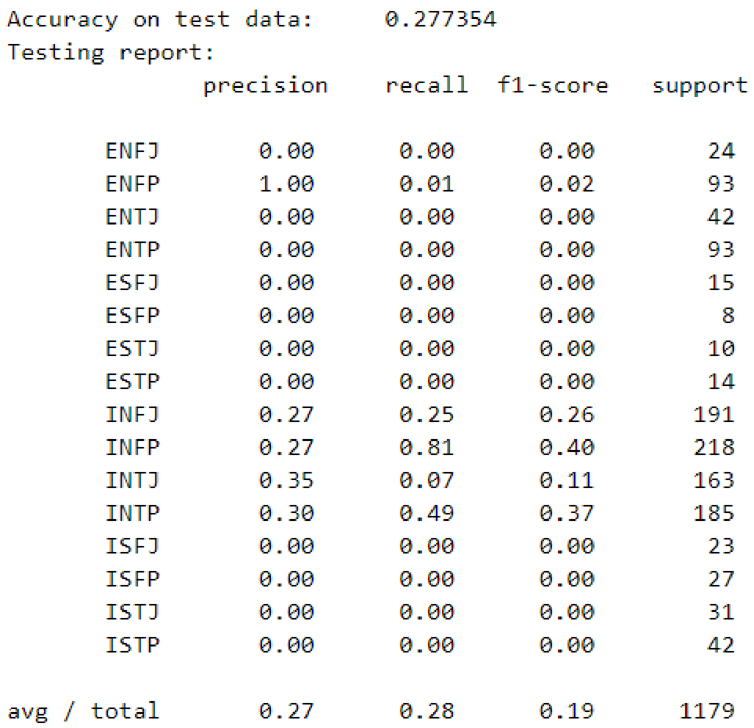
The Kaggle dataset's bias is show above with the clear variation between the introverted and extroverted types. The reddit database's distribution is shown below, it is biased as well but also helps to raise the number of examples of the uncommonly posting types.



The combined samples with some extra Reddit are shown below.

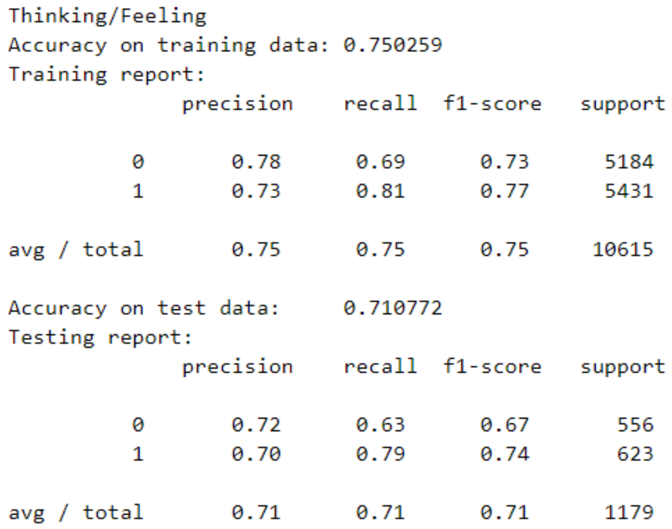


To clean the text data I removed English stop words, replaced URLs with <URL> and replaced all the 4 letter personality types with <MBTI>. For the existing Kaggle dataset sample the <MBTI> replacement factored in at about 20% of some of the models, so eliminating that means there may need more data or better initial text analysis. The accuracy general accuracy of predicting the test type from the original data without the Myers Brigg class in the text is shown below.



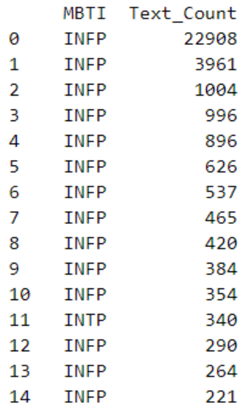
```
Accuracy on test data:     0.277354
Testing report:
             precision    recall  f1-score   support

       ENFJ       0.00      0.00      0.00        24
       ENFP       1.00      0.01      0.02        93
       ENTJ       0.00      0.00      0.00        42
       ENTP       0.00      0.00      0.00        93
       ESFJ       0.00      0.00      0.00        15
       ESFP       0.00      0.00      0.00         8
       ESTJ       0.00      0.00      0.00        10
       ESTP       0.00      0.00      0.00        14
       INFJ       0.27      0.25      0.26       191
       INFP       0.27      0.81      0.40       218
       INTJ       0.35      0.07      0.11       163
       INTP       0.30      0.49      0.37       185
       ISFJ       0.00      0.00      0.00        23
       ISFP       0.00      0.00      0.00        27
       ISTJ       0.00      0.00      0.00        31
       ISTP       0.00      0.00      0.00        42

avg / total       0.27      0.28      0.19      1179
```

It is clear that the model is not currently very accurate, but for the Thinking/Feeling (T/F) split the predictions are reasonably accurate.



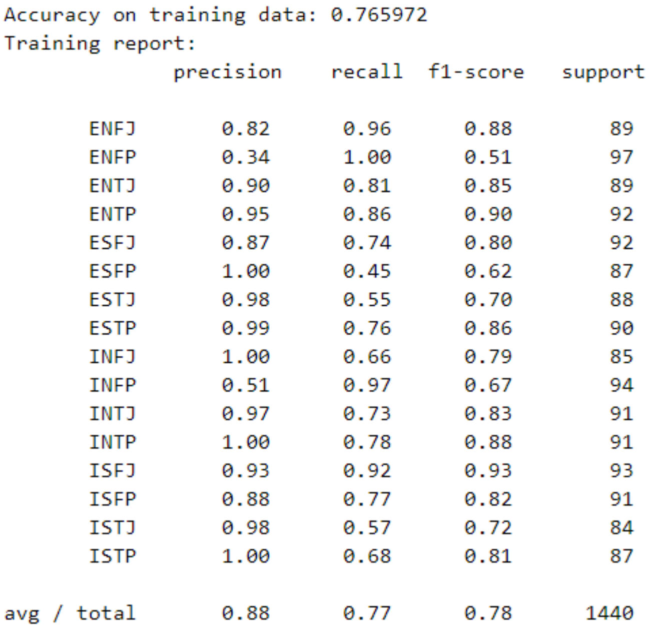
```
Thinking/Feeling
Accuracy on training data: 0.750259
Training report:
             precision    recall  f1-score   support

          0       0.78      0.69      0.73      5184
          1       0.73      0.81      0.77      5431

avg / total       0.75      0.75      0.75     10615

Accuracy on test data:     0.710772
Testing report:
             precision    recall  f1-score   support

          0       0.72      0.63      0.67       556
          1       0.70      0.79      0.74       623

avg / total       0.71      0.71      0.71      1179
```

I also tested the accuracy of predicting my and a couple close friend's personality types from my texting history and compared them to their stated MBTI. It predicts INFP no matter what, so I think the unbalanced nature of the set is skewing the results towards the most common.



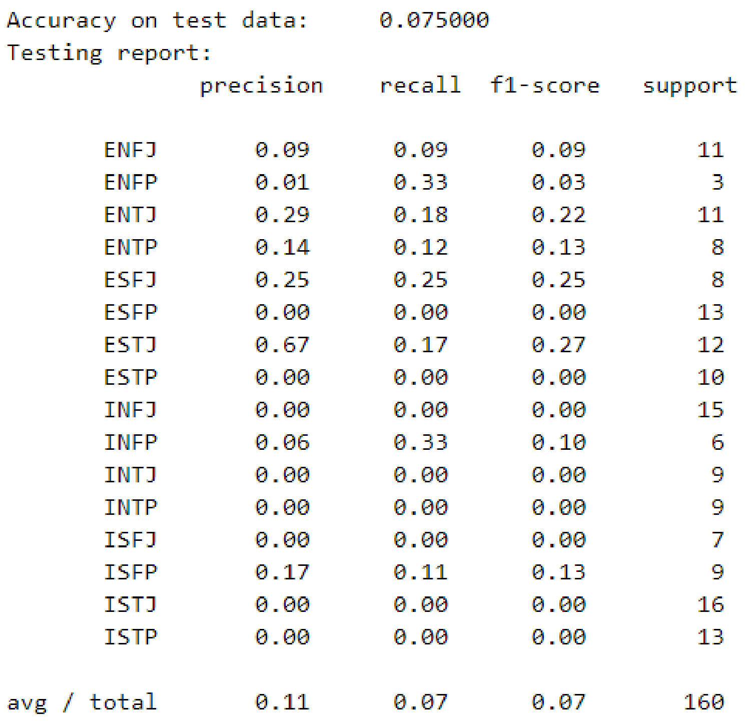
```
    MBTI  Text_Count
0   INFP       22908
1   INFP        3961
2   INFP        1004
3   INFP         996
4   INFP         896
5   INFP         626
6   INFP         537
7   INFP         465
8   INFP         420
9   INFP         384
10  INFP         354
11  INTP         340
12  INFP         290
13  INFP         264
14  INFP         221
```

When I only use 100 different users from each type it is more accurate in predicting each individual type, but was less likely to get them all right at once (presumably since it was no longer right by just predicting the most common thus eliminating the possible assignments to about 4-6 categories). First the categorization of all classes at once is shown with 1600 balanced users, then the categorization of each type individually.



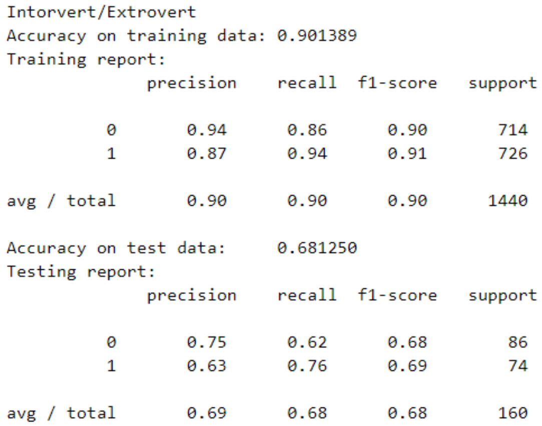
```
Accuracy on training data: 0.765972
Training report:
             precision    recall  f1-score   support

       ENFJ       0.82      0.96      0.88        89
       ENFP       0.34      1.00      0.51        97
       ENTJ       0.90      0.81      0.85        89
       ENTP       0.95      0.86      0.90        92
       ESFJ       0.87      0.74      0.80        92
       ESFP       1.00      0.45      0.62        87
       ESTJ       0.98      0.55      0.70        88
       ESTP       0.99      0.76      0.86        90
       INFJ       1.00      0.66      0.79        85
       INFP       0.51      0.97      0.67        94
       INTJ       0.97      0.73      0.83        91
       INTP       1.00      0.78      0.88        91
       ISFJ       0.93      0.92      0.93        93
       ISFP       0.88      0.77      0.82        91
       ISTJ       0.98      0.57      0.72        84
       ISTP       1.00      0.68      0.81        87

avg / total       0.88      0.77      0.78      1440
```

The accuracy on the test data to each personality type is low for the smaller sample size of 100 users per type.

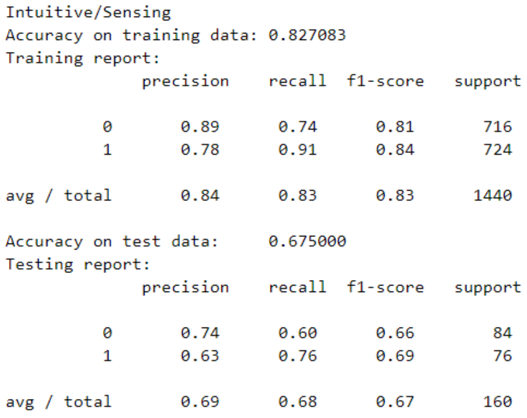


```
Accuracy on test data:     0.075000
Testing report:
              precision    recall  f1-score   support

       ENFJ       0.09      0.09      0.09        11
       ENFP       0.01      0.33      0.03         3
       ENTJ       0.29      0.18      0.22        11
       ENTP       0.14      0.12      0.13         8
       ESFJ       0.25      0.25      0.25         8
       ESFP       0.00      0.00      0.00        13
       ESTJ       0.67      0.17      0.27        12
       ESTP       0.00      0.00      0.00        10
       INFJ       0.00      0.00      0.00        15
       INFP       0.06      0.33      0.10         6
       INTJ       0.00      0.00      0.00         9
       INTP       0.00      0.00      0.00         9
       ISFJ       0.00      0.00      0.00         7
       ISFP       0.17      0.11      0.13         9
       ISTJ       0.00      0.00      0.00        16
       ISTP       0.00      0.00      0.00        13

avg / total       0.11      0.07      0.07       160
```

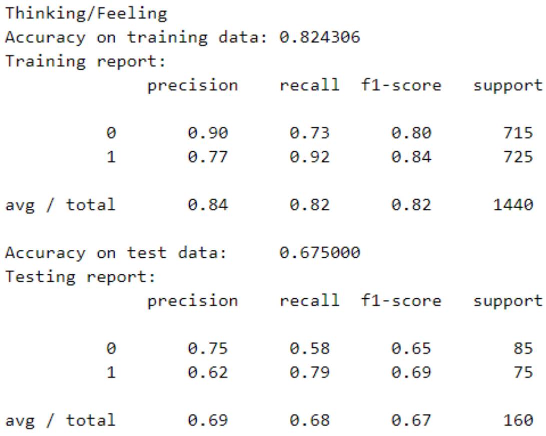
Each estimation individually is fairly on its own, just all 4 are hard to get right at the same time.



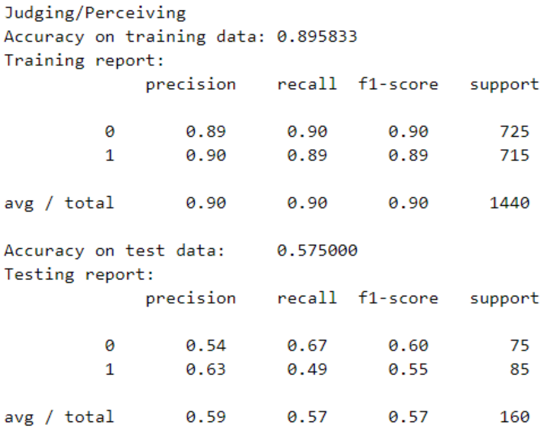
```
Intorvert/Extrovert
Accuracy on training data: 0.901389
Training report:
             precision    recall  f1-score   support

          0       0.94      0.86      0.90       714
          1       0.87      0.94      0.91       726

avg / total       0.90      0.90      0.90      1440

Accuracy on test data:     0.681250
Testing report:
             precision    recall  f1-score   support

          0       0.75      0.62      0.68        86
          1       0.63      0.76      0.69        74

avg / total       0.69      0.68      0.68       160
```



```
Intuitive/Sensing
Accuracy on training data: 0.827083
Training report:
             precision    recall  f1-score   support

          0       0.89      0.74      0.81       716
          1       0.78      0.91      0.84       724

avg / total       0.84      0.83      0.83      1440

Accuracy on test data:     0.675000
Testing report:
             precision    recall  f1-score   support

          0       0.74      0.60      0.66        84
          1       0.63      0.76      0.69        76

avg / total       0.69      0.68      0.67       160
```

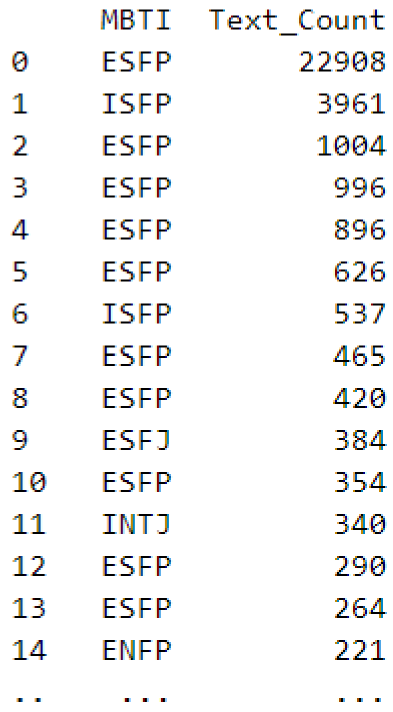


```
Thinking/Feeling
Accuracy on training data: 0.824306
Training report:
             precision    recall  f1-score   support

          0       0.90      0.73      0.80       715
          1       0.77      0.92      0.84       725

avg / total       0.84      0.82      0.82      1440

Accuracy on test data:     0.675000
Testing report:
             precision    recall  f1-score   support

          0       0.75      0.58      0.65        85
          1       0.62      0.79      0.69        75

avg / total       0.69      0.68      0.67       160
```



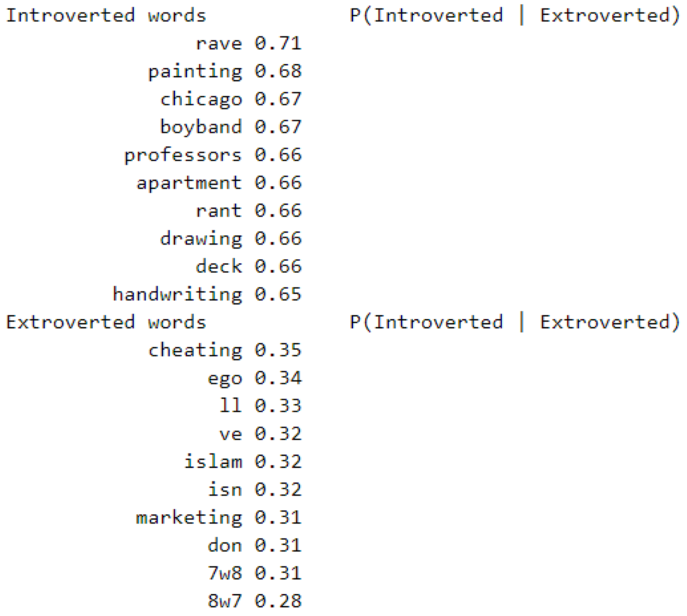
```
Judging/Perceiving
Accuracy on training data: 0.895833
Training report:
             precision    recall  f1-score   support

          0       0.89      0.90      0.90       725
          1       0.90      0.89      0.89       715

avg / total       0.90      0.90      0.90      1440

Accuracy on test data:     0.575000
Testing report:
             precision    recall  f1-score   support

          0       0.54      0.67      0.60        75
          1       0.63      0.49      0.55        85

avg / total       0.59      0.57      0.57       160
```

The types assigned based on texting history vary a lot but none of the 4 people I know have an accurate assignment.



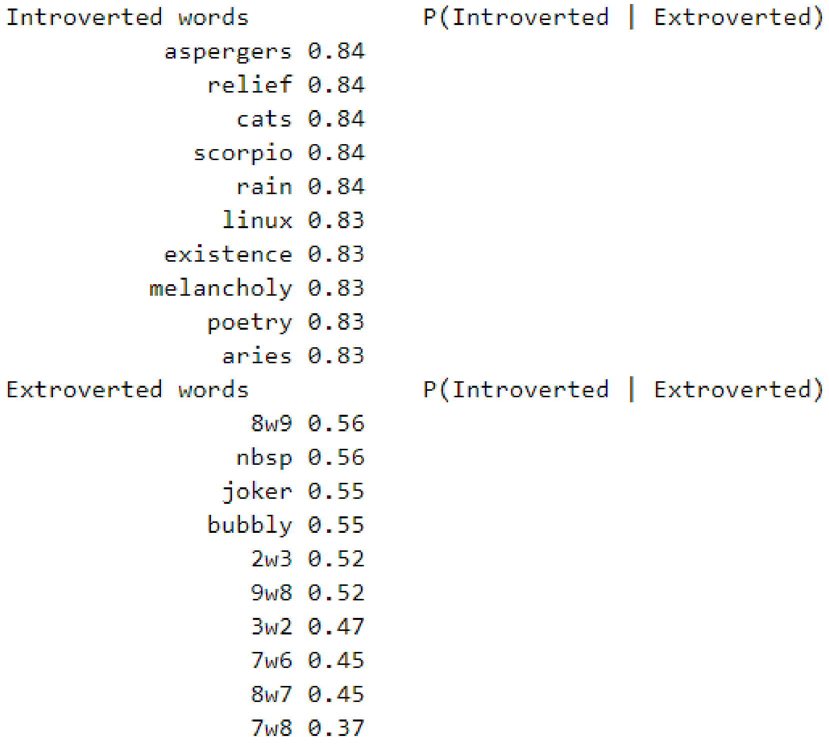
```
    MBTI  Text_Count
0   ESFP       22908
1   ISFP        3961
2   ESFP        1004
3   ESFP         996
4   ESFP         896
5   ESFP         626
6   ISFP         537
7   ESFP         465
8   ESFP         420
9   ESFJ         384
10  ESFP         354
11  INTJ         340
12  ESFP         290
13  ESFP         264
14  ENFP         221
..   ...         ...
```

The introverted words don't seem as accurate in general, such as 'rave' seems more extroverted to me, but the Extroverted words including 'ego' seem better than before.



```
Introverted words            P(Introverted | Extroverted)
                rave 0.71
            painting 0.68
             chicago 0.67
             boyband 0.67
          professors 0.66
           apartment 0.66
                rant 0.66
             drawing 0.66
                deck 0.66
         handwriting 0.65
Extroverted words            P(Introverted | Extroverted)
            cheating 0.35
                 ego 0.34
                  ll 0.33
                  ve 0.32
               islam 0.32
                 isn 0.32
           marketing 0.31
                 don 0.31
                 7w8 0.31
                 8w7 0.28
```

For comparison the introverted and extroverted words for all the users are shown below. The introverted word seem more accurate, but all the actual words that are rated extroverted are still above .5 (which still indicates introversion).



```
Introverted words            P(Introverted | Extroverted)
           aspergers 0.84
              relief 0.84
                cats 0.84
             scorpio 0.84
                rain 0.84
               linux 0.83
           existence 0.83
          melancholy 0.83
              poetry 0.83
               aries 0.83
Extroverted words            P(Introverted | Extroverted)
                 8w9 0.56
                nbsp 0.56
               joker 0.55
              bubbly 0.55
                 2w3 0.52
                 9w8 0.52
                 3w2 0.47
                 7w6 0.45
                 8w7 0.45
                 7w8 0.37
```