Parker Williamson
5/9/2018
Springboard Data Science Career Track

Capstone Project 2 Final Report – 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.
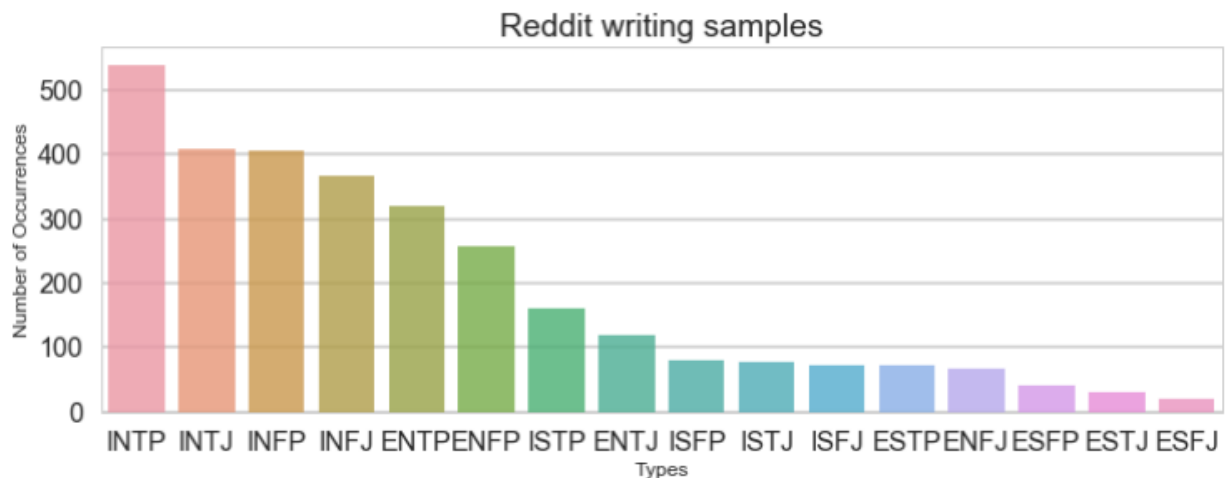
Dataset

I focused on accurately predicting MBTI personality types. The MBTI raw data I will get from Kaggle's Myers Briggs dataset (https://www.kaggle.com/datasnaek/mbti-type). Some personality types didn'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 collected from Reddit's MBTI subreddits which have posters self labelled type and separate subreddits for different types. Text data was normalized using (http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/).

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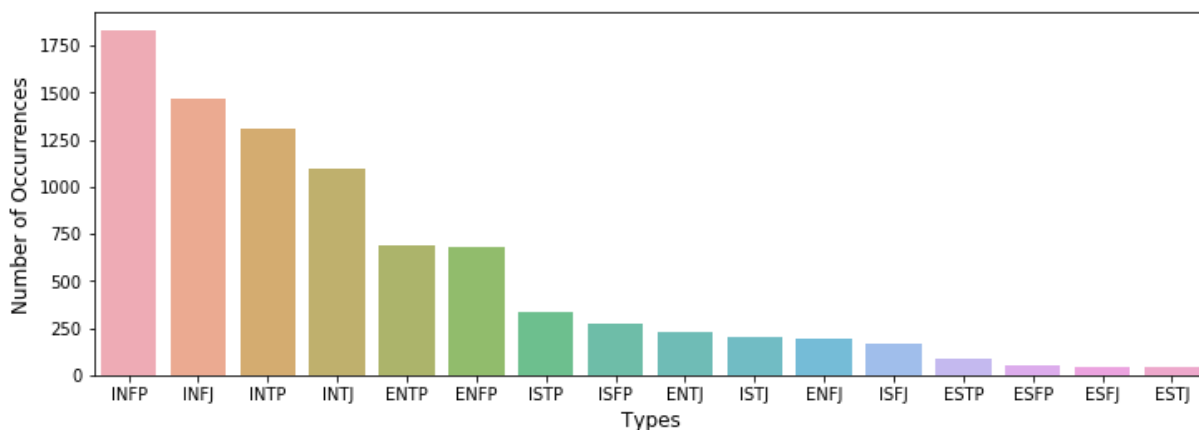
The Kaggle dataset's bias is show above with the clear variation between the introverted and extroverted types. The reddit database's distribution is shown below, it is biased as well but also helps to raise the number of examples of the uncommonly posting types.

Reddit writing samples

The combined samples with some extra Reddit are shown below.

Combined writing samples
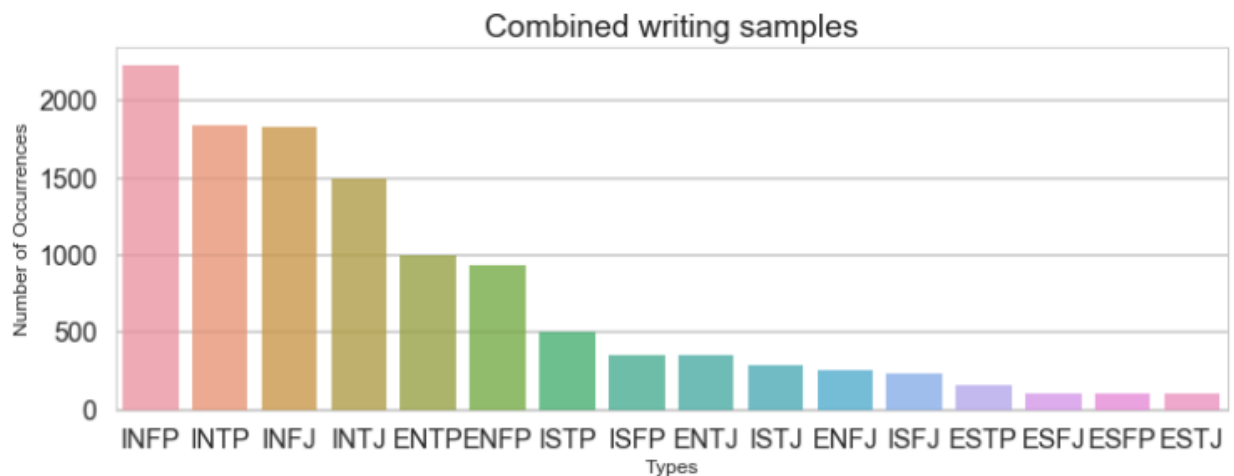
To clean the text data I removed English stop words, replaced URLs with <URL> and replaced all the 4 letter personality types with <MBTI>. For the existing Kaggle dataset sample the <MBTI> replacement factored in at about 20% of some of the models, so eliminating that means there may need more data or 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 good at predicting all the personality traits at once. The predictions for each individual traits are 60-70% accurate depending on the training test split.

```
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. The difference between texting and forum posts made the predictions all the same until I normalized the results to a dataset of texting I found online.

```
       MBTI  Text_Count
0      ISTJ       23055
1      ESTJ        3957
2      INTJ        1003
3      INTP         989
4      INTJ         893
5      ENFP         636
6      ESFP         537
7      ESTP         508
8      INTJ         415
9      ENFJ         384
10     ISTJ         354
11     ENFP         340
12     INTJ         290
13     ISTJ         265
14     ISTJ         221
```

When I only use 100 different users from each type there was a lot more variance and although the training accuracy was greater the testing accuracy was very unhelpful. It had been estimating the most likely and there was not enough new info to accurately estimate using.

```
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 around 65% on its own, just all 4 are hard to get right at the same time.

```
Intorvert/Extrovert                         Intuitive/Sensing
Accuracy on training data: 0.738052         Accuracy on training data: 0.758993
Training report:                            Training report:
          precision  recall  f1-score  support           precision  recall  f1-score  support

      0      0.49     0.82     0.61     1986         0     0.37     0.75     0.49     1210
      1      0.92     0.71     0.80     5798         1     0.94     0.76     0.84     6574

avg / total  0.81     0.74     0.75     7784   avg / total  0.85     0.76     0.79     7784

Accuracy on test data:     0.675062         Accuracy on test data:     0.695511
Testing report:                             Testing report:
          precision  recall  f1-score  support           precision  recall  f1-score  support

      0      0.42     0.67     0.51     1023         0     0.26     0.52     0.35      623
      1      0.86     0.68     0.76     2987         1     0.89     0.73     0.80     3387

avg / total  0.74     0.68     0.69     4010   avg / total  0.79     0.70     0.73     4010
```

```
Thinking/Feeling                                 Judging/Perceiving
Accuracy on training data: 0.740365              Accuracy on training data: 0.707862
Training report:                                 Training report:
           precision   recall  f1-score  support            precision   recall  f1-score  support

        0       0.78     0.65      0.71     3789          0       0.79     0.71      0.74     4705
        1       0.71     0.83      0.77     3995          1       0.61     0.71      0.66     3079

avg / total     0.75     0.74      0.74     7784 avg / total     0.72     0.71      0.71     7784

Accuracy on test data:     0.718703              Accuracy on test data:     0.628678
Testing report:                                  Testing report:
           precision   recall  f1-score  support            precision   recall  f1-score  support

        0       0.77     0.61      0.68     1951          0       0.71     0.64      0.68     2424
        1       0.69     0.83      0.75     2059          1       0.53     0.61      0.56     1586

avg / total     0.73     0.72      0.71     4010 avg / total     0.64     0.63      0.63     4010
```

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	ISTJ	23055
1	ESTJ	3957
2	INTJ	1003
3	INTP	989
4	INTJ	893
5	ENFP	636
6	ESFP	537
7	ESTP	508
8	INTJ	415
9	ENFJ	384
10	ISTJ	354
11	ENFP	340
12	INTJ	290
13	ISTJ	265
14	ISTJ	221

The introverted words do tend to make sense as more intellectual or emotional.

```
Introverted words          P(Introverted | Extroverted)
           linux 0.64
           lucid 0.64
         fantasy 0.62
           genre 0.62
            neat 0.62
             sky 0.62
          poetry 0.62
          genres 0.61
          relief 0.61
        daydream 0.61
            fond 0.61
        movement 0.60
            anna 0.60
            rain 0.60
           stare 0.60
       existence 0.60
              64 0.60
         adopted 0.60
        ordinary 0.60
         possess 0.60
Extroverted words          P(Introverted | Extroverted)
             nts 0.40
           fucks 0.40
            awww 0.40
              du 0.40
            yall 0.40
             omg 0.40
           stark 0.40
              xd 0.40
        charming 0.40
              bc 0.40
            didn 0.40
         subtype 0.40
           charm 0.40
             det 0.40
             mod 0.40
         sensors 0.40
          racist 0.39
             owo 0.39
              en 0.39
             ben 0.39
```

Overall it is clear that the MLP classifier is an okay, but not accurate enough model for personality prediction based on text. It is around 65-70% accurate per Meyers Briggs personality trait. To be in line with the normal method of testing personality it would need to be around 85% accurate. This does show how Reddit can be used to expand the dataset some, and how different types of posts tend to indicate different personality types.