Capstone Project 2 Final Report
Text Mash

PARKER WILLIAMSON

5/9/2018

SPRINGBOARD DATA SCIENCE CAREER TRACK

Problem

- Use text to assign Meyers Briggs personality types in order to improve interactions between people and improve user experience

Client

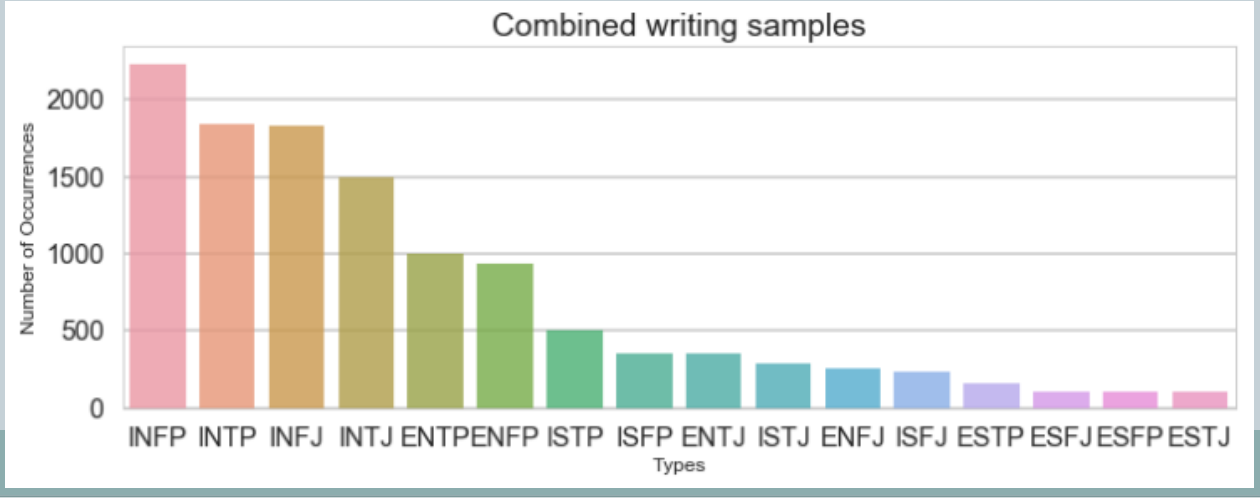
- Apps could use text history to make recommendations based on personality type

Datasets

- The main dataset isKaggle's Myers Briggs dataset (https://www.kaggle.com/datasnaek/mbti-type)
- Reddit data from (https://www.reddit.com/r/mbti/) was also collected using the reddit API PRAW
- Text data was normalized using (http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/).

Findings

- The personality datasets are very biased, because personality types such as introverts are much more likely to post
- The Reddit and Kaggle dataset distribution combined is shown below



Findings

- 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.

Findings

- I tested stemming as well but its effect on the results was negligible (two of the types went up and two went down by nearly the same amount)
  - Since stemming is also slower I excluded it from my models after this test

Findings

- Predicting all the personality types at once is not possible, because the bias of the dataset overwrites the information extracted from the text.

Accuracy on test data: 0.261845

Testing report:

	precision	recall	f1-score	support
ENFJ	0.00	0.00	0.00	87
ENFP	0.00	0.00	0.00	317
ENTJ	0.00	0.00	0.00	119
ENTP	0.00	0.00	0.00	342
ESFJ	0.00	0.00	0.00	35
ESFP	0.00	0.00	0.00	35
ESTJ	0.00	0.00	0.00	34
ESTP	0.00	0.00	0.00	54
INFJ	0.25	0.22	0.24	624
INFP	0.25	0.83	0.38	761
INTJ	0.36	0.04	0.06	510
INTP	0.30	0.42	0.35	627
ISFJ	0.00	0.00	0.00	81
ISFP	0.00	0.00	0.00	119
ISTJ	0.00	0.00	0.00	96
ISTP	0.00	0.00	0.00	169
avg / total	0.18	0.26	0.17	4010

Findings

- Numerous different models were tested:
  - MultinomialNB
  - LinearSVC
  - LogisticRegression
  - SGDClassifier
  - XGBClassifier
  - CNN
  - RNN
- The different classifiers varied from ~63%-73% on IE
- They were similar for the other types
- In order of smallest F1 score: LogReg, MultNB, SGD, XGboost, LinearSVC

Findings

- The test accuracy of the personality types are ~65-70% using MultinomialNB

Intorvert/Extrovert
Accuracy on training data: 0.738052
Training report:

	precision	recall	f1-score	support
0	0.49	0.82	0.61	1986
1	0.92	0.71	0.80	5798
avg / total	0.81	0.74	0.75	7784

Accuracy on test data: 0.675062
Testing report:

	precision	recall	f1-score	support
0	0.42	0.67	0.51	1023
1	0.86	0.68	0.76	2987
avg / total	0.74	0.68	0.69	4010

Intuitive/Sensing
Accuracy on training data: 0.758993
Training report:

	precision	recall	f1-score	support
0	0.37	0.75	0.49	1210
1	0.94	0.76	0.84	6574
avg / total	0.85	0.76	0.79	7784

Accuracy on test data: 0.695511
Testing report:

	precision	recall	f1-score	support
0	0.26	0.52	0.35	623
1	0.89	0.73	0.80	3387
avg / total	0.79	0.70	0.73	4010

Thinking/Feeling
Accuracy on training data: 0.740365
Training report:

	precision	recall	f1-score	support
0	0.78	0.65	0.71	3789
1	0.71	0.83	0.77	3995
avg / total	0.75	0.74	0.74	7784

Accuracy on test data: 0.718703
Testing report:

	precision	recall	f1-score	support
0	0.77	0.61	0.68	1951
1	0.69	0.83	0.75	2059
avg / total	0.73	0.72	0.71	4010

Judging/Perceiving
Accuracy on training data: 0.707862
Training report:

	precision	recall	f1-score	support
0	0.79	0.71	0.74	4705
1	0.61	0.71	0.66	3079
avg / total	0.72	0.71	0.71	7784

Accuracy on test data: 0.628678
Testing report:

	precision	recall	f1-score	support
0	0.71	0.64	0.68	2424
1	0.53	0.61	0.56	1586
avg / total	0.64	0.63	0.63	4010

Adding the Reddit Data

- Including the reddit data decreased the test accuracy by about 2%, but the Kaggle MBTI data was not as applicable to average texting, because it is based on text where people are talking about their personalities. Adding the Reddit data helps prevent overfitting, for example, one of the most strongly weighted words (towards extroversion) is socionics. Socionics is not a word used in many texting conversations, but it was used fairly commonly in the MBTI forum drawn from for the Kaggle dataset.

- The word that are the most polarized

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

Findings

- There is clear variance in assigned personality types once the texts had been normalized

	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