What are you saying?

Introversion/Extroversion Algorithm: Mod 5 Project

Alaska Lam

Problem Statement:

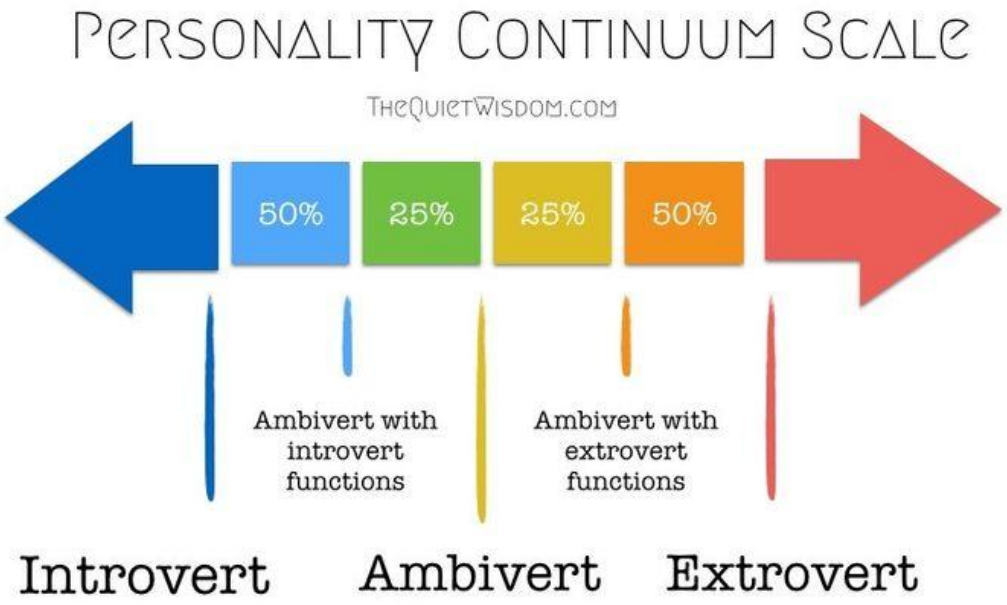
To better understand possible differences in online writing style between introverts and extroverts

*writing conversationally about themselves.

Business Value:

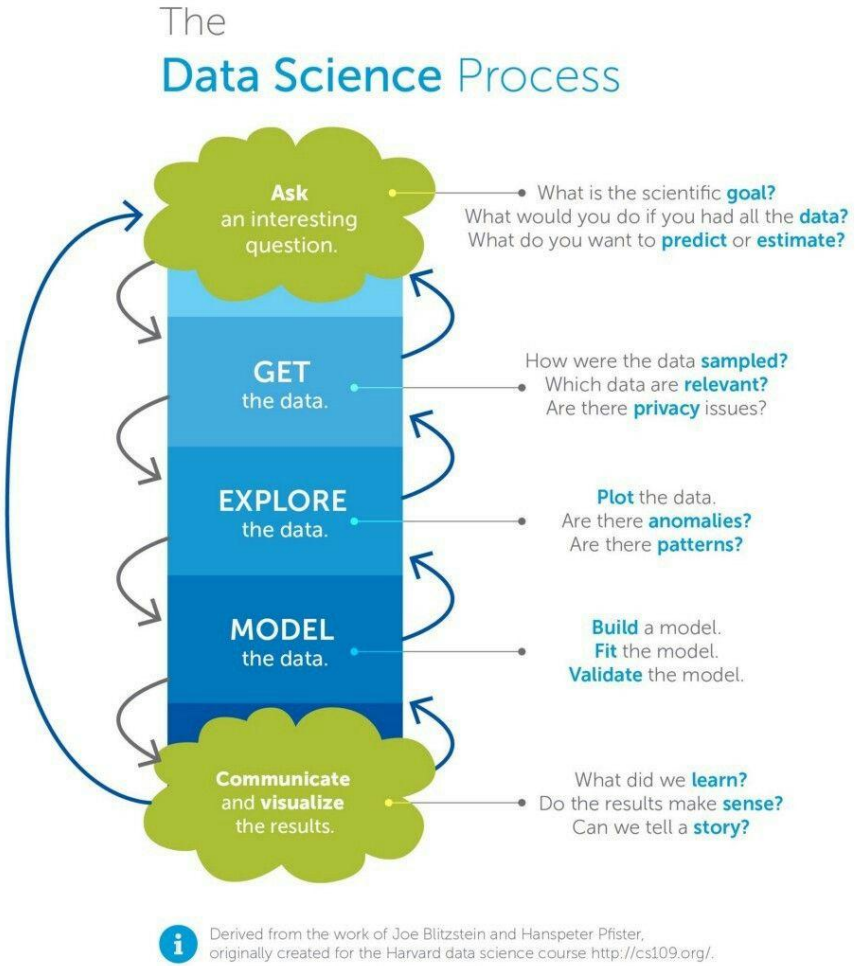
Create foundation for more comprehensive friendship-algorithm

Capitalize on marketing demographics



Methodology:

Analyze past text-data written on Personality Cafe.net, a forum for talking about your life experiences



Model Results - Focus on F1 Score: Introversion vs Extroversion

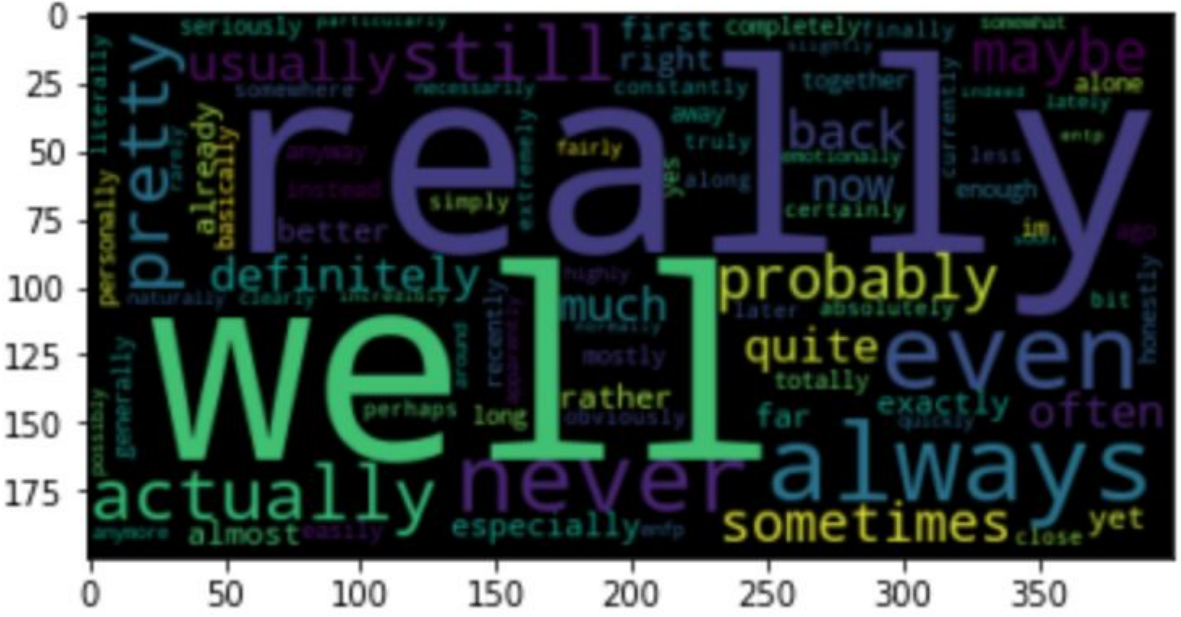
F1 Score = Reduce General Mislabelling Errors

*Weighted average

```
Multinomial Naive Bayes
Training Accuracy: 0.9283                    Testing Accuracy: 0.6923

[[1531  693]
 [ 188  451]]
0.6922808243101641
1531 693 188 451
Classification Matrix:
              precision    recall  f1-score   support

           0       0.89      0.69      0.78      2224
           1       0.39      0.71      0.51       639

    accuracy                           0.69      2863
   macro avg       0.64      0.70      0.64      2863
weighted avg       0.78      0.69      0.72      2863
```

Recommendation 1, part 1:

Introvert vs extrovert word clouds are similar - instead look at combinations of words and phrases



Recommendation 1, part 2:

Look at bigrams -

“!!” is double as common in extroverts

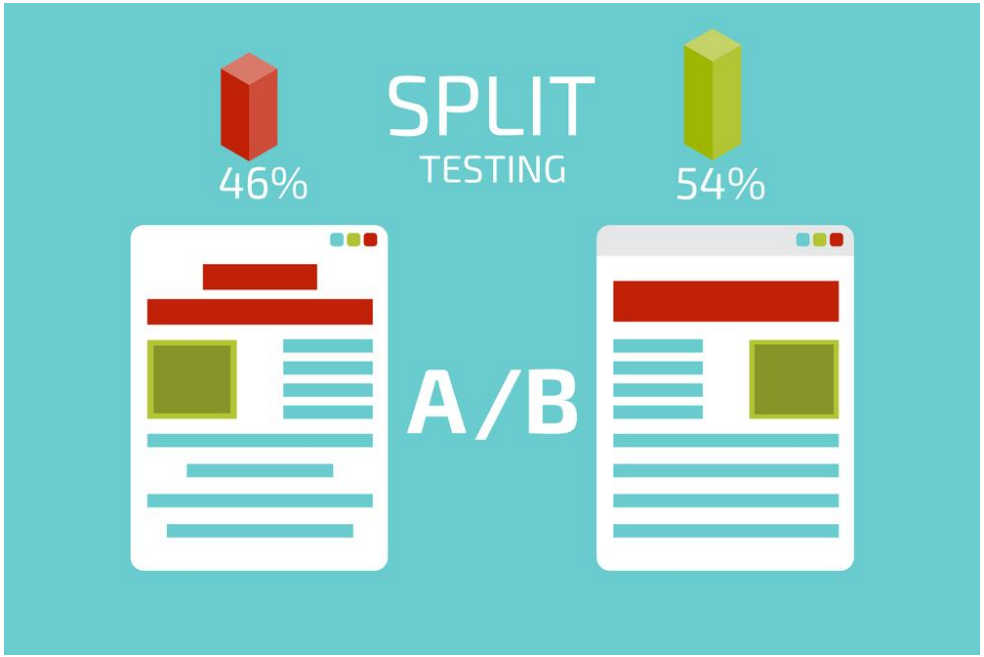
“me!” 1.5x for extroverts

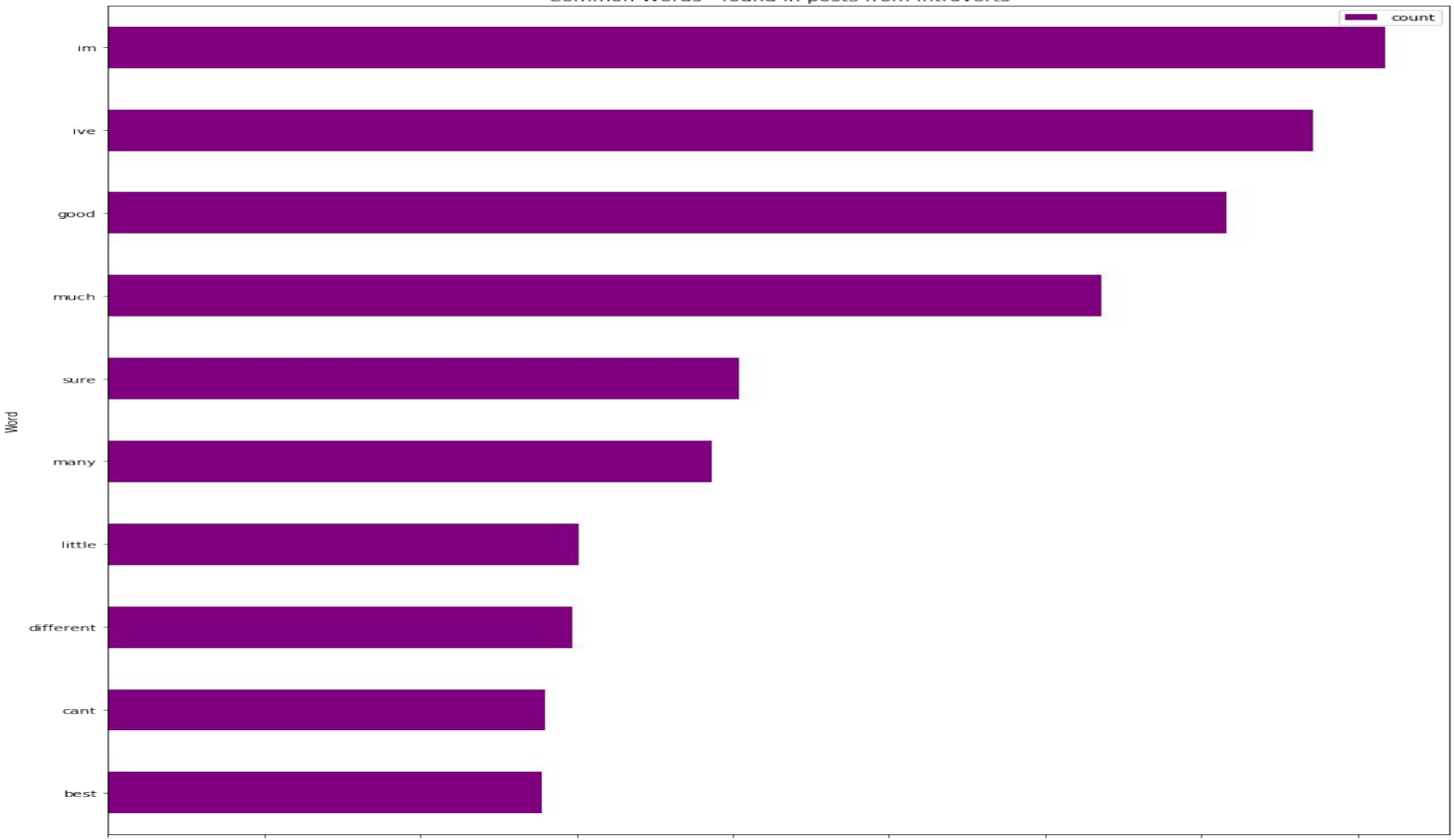
“too!” high on extroverts’ list

Recommendation 2:

Since all top words are similar when comparing introverts vs extroverts, do not attempt to create separate marketing campaigns at this point -

perhaps later for domain -specific algorithms




Common Words - found in posts from introverts
count
Word
im
ive
good
much
sure
many
little
different
cant
best
0
2000
4000
6000
8000
10000
12000
14000
16000

Recommendation 3:

Mutual information scores – correlation with literary references in introverts

Introverts - phrases that do not show up in extroverts' top pmi

giga blender - top one

whats subtype

monte cristo

jrr tolkien

forrest gump

los angeles

englishlanguage, arts

sodium,sodium

mockingbird,harper

bungee jumpingskydiving

dalai lama

winnie pooh

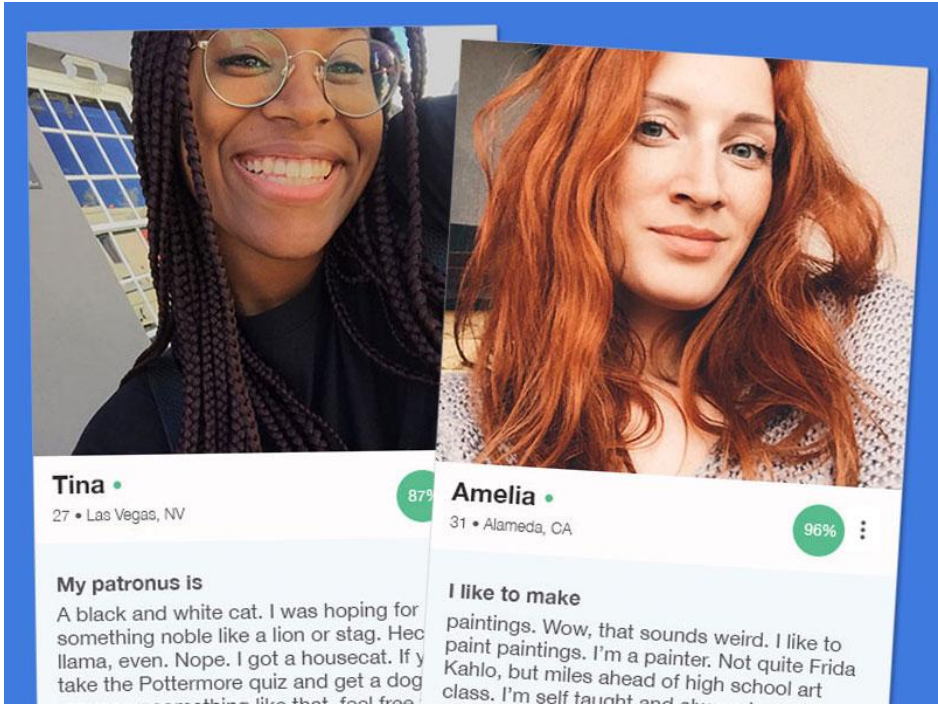
Future Considerations

Continue to fine-tune the current models, reduce overfitting

New model based on how often top bigrams or bigrams with top mutual info scores are used

Obtain more text-data from social events and websites such as OkCupid

Incorporate into friendship algorithm



Thank You!