# Introvert or Extrovert:
# A Myers-Briggs Experiment

Jose Reyes

02/01/2018

# Introduction

# Human Personality

- For ages scholars have studied & aimed to classify human personality.

- Various approaches:

1. Humorism: Ancient Greeks used the balance of body fluids to classify personality.

2. Phrenology/Physiognomy: Classify personality based on outward appearance.

3. Psychological: Myers-Briggs, Rorschach, and Thematic Apperception Test (TAT).

# Assessing Personality

- Reasons why:
  - Form an accurate picture of an individual.
  - If personality is know, easy to predict future actions.
  - Easy to identify preferred preferences.

- Problems:
  - Time-consuming
  - Issues of low-reliability
  - Issues of low-validity

# The Solution

- If we can analyze social media posts to assess personality we can:

  1. Identify key words/speech patterns that determine personality type.
  2. Firms can use findings to better target marketing campaigns.
  3. Education institutions can better tailor their lessons to students of different personality types.

- Using Python and the Myers-Briggs Personality Indicator dataset from the Kaggle website, we will aim to classify people as Introverts or Extroverts using text data.

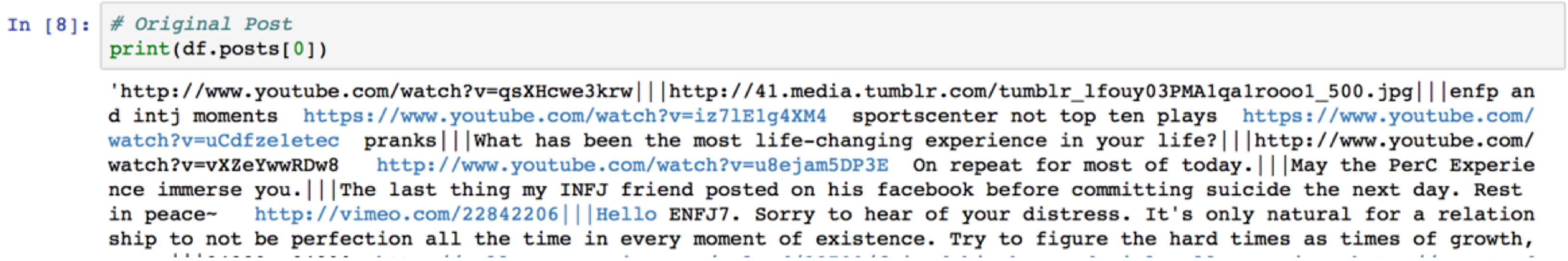
# The Dataset

# Dataset Inspection

- Dataset contains 2 columns:
  - A collection of a user’s internet posts, concatenated as a string.
  - The user’s Myers-Briggs Personality Identifier (MBTI).

- 8,675 Observations

- Sample posts:



```
In [8]: # Original Post
        print(df.posts[0])
```

```
'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp an
d intj moments  https://www.youtube.com/watch?v=iz7lE1g4XM4  sportscenter not top ten plays  https://www.youtube.com/
watch?v=uCdfze1etec  pranks|||What has been the most life-changing experience in your life?|||http://www.youtube.com/
watch?v=vXZeYwwRDw8   http://www.youtube.com/watch?v=u8ejam5DP3E  On repeat for most of today.|||May the PerC Experie
nce immerse you.|||The last thing my INFJ friend posted on his facebook before committing suicide the next day. Rest
in peace~   http://vimeo.com/22842206|||Hello ENFJ7. Sorry to hear of your distress. It's only natural for a relation
ship to not be perfection all the time in every moment of existence. Try to figure the hard times as times of growth,
```

# Data Wrangling

# Assumptions

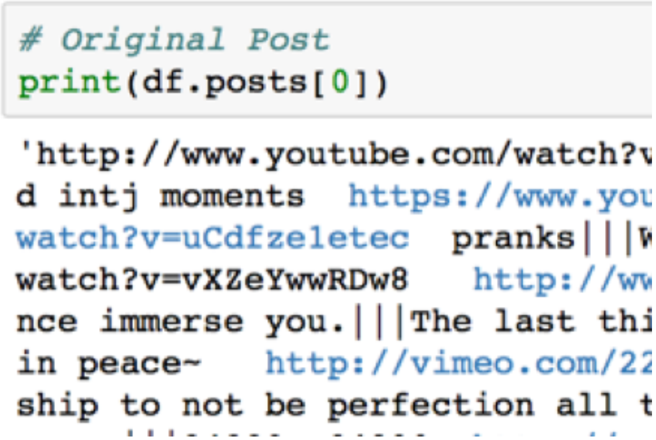
- Each observation is a unique observation, no unique identifier.
- All user posts are truly self-reported, no high jacked accounts.
- All user posts were posted over time, not all posts in one day.

# Initial Data Quality Findings

- Relatively clean dataset with minor exceptions:
  - Posts separated by a triple pipe ("|||")
  - Posts contain URLs, numeric data, and MBTI codes.
  - Posts contain a multiple of stop words (i.e. "the", "of", "a", "though" … etc.)

# Data Wrangling

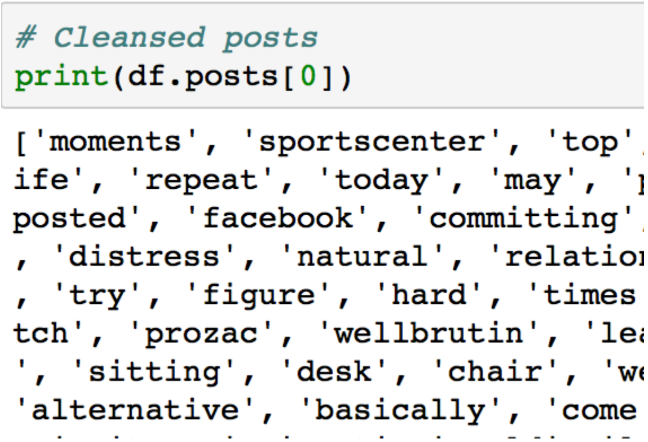
**Original Posts (Sample)**



```
# Original Post
print(df.posts[0])
```

```
'http://www.youtube.com/watch?v
d intj moments  https://www.you
watch?v=uCdfze1etec  pranks|||W
watch?v=vXZeYwwRDw8   http://ww
nce immerse you.|||The last thi
in peace~   http://vimeo.com/22
ship to not be perfection all t
```

**Cleansed Posts (Sample)**



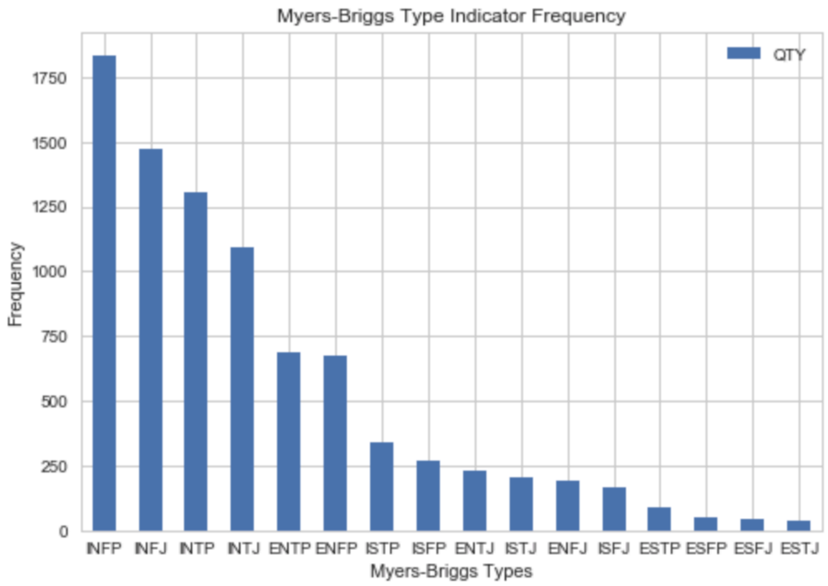
```
# Cleansed posts
print(df.posts[0])
```

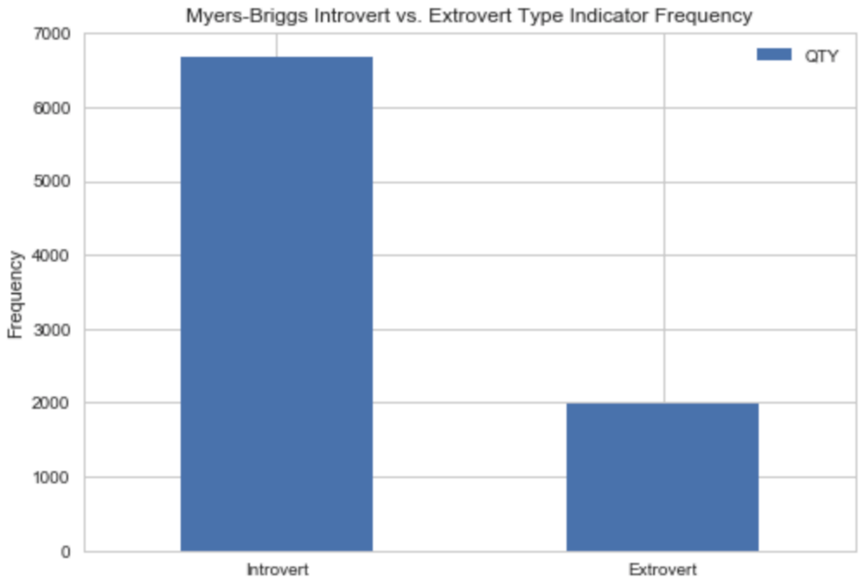
```
['moments', 'sportscenter', 'top'
ife', 'repeat', 'today', 'may', '
posted', 'facebook', 'committing'
, 'distress', 'natural', 'relatio
, 'try', 'figure', 'hard', 'times
tch', 'prozac', 'wellbrutin', 'lea
', 'sitting', 'desk', 'chair', 'we
'alternative', 'basically', 'come
```

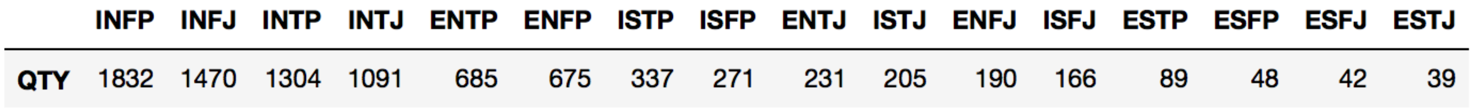
# Data Wrangling (con’t)

- Additional data wrangling includes value transformation:
  - Creation of a 3rd column: “**Attitude**”
    - Denotes if user is Introvert or Extrovert, based on MBTI value. If MBTI starts with “I,” user is Introvert. If MBTI starts with “E,” user is Extrovert.
  - Creation of a 4th column: “**word_qty**”
    - Denotes word count of the cleansed string.

# Inferential Statistics



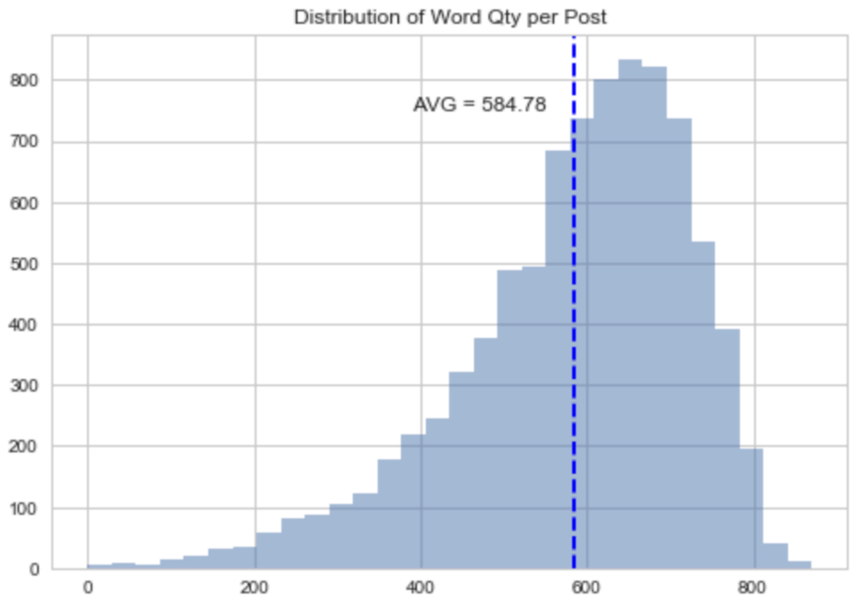




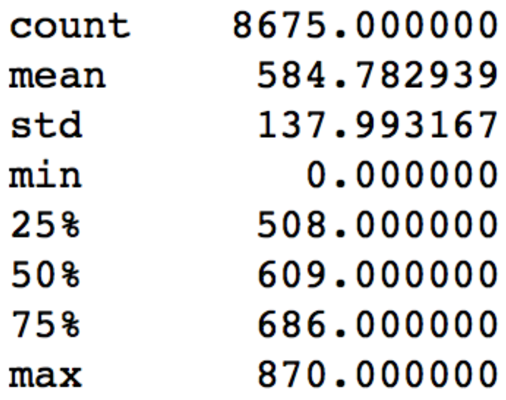
| | INFP | INFJ | INTP | INTJ | ENTP | ENFP | ISTP | ISFP | ENTJ | ISTJ | ENFJ | ISFJ | ESTP | ESFP | ESFJ | ESTJ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **QTY** | 1832 | 1470 | 1304 | 1091 | 685 | 675 | 337 | 271 | 231 | 205 | 190 | 166 | 89 | 48 | 42 | 39 |

# Inferential Statistics (con’t)

## String Word Count Distribution



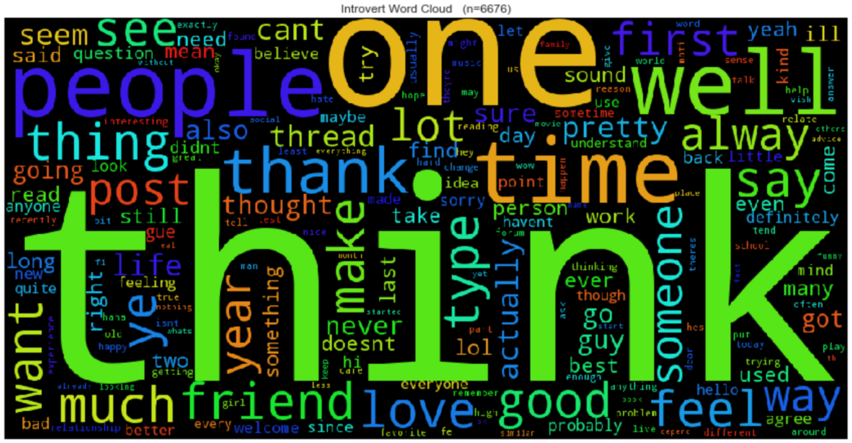
## String Word Count Descriptive Statistics



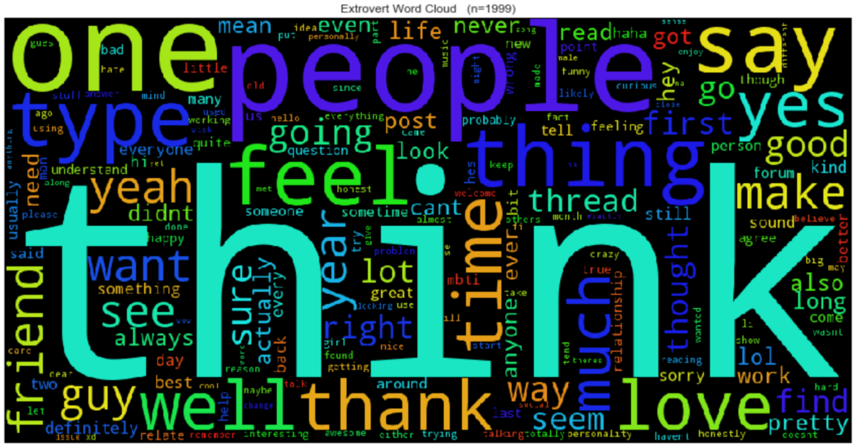
```
count     8675.000000
mean       584.782939
std        137.993167
min          0.000000
25%        508.000000
50%        609.000000
75%        686.000000
max        870.000000
```

# Inferential Statistics (con't)

**Word Cloud of Introvert posts:**
*n=6,676*



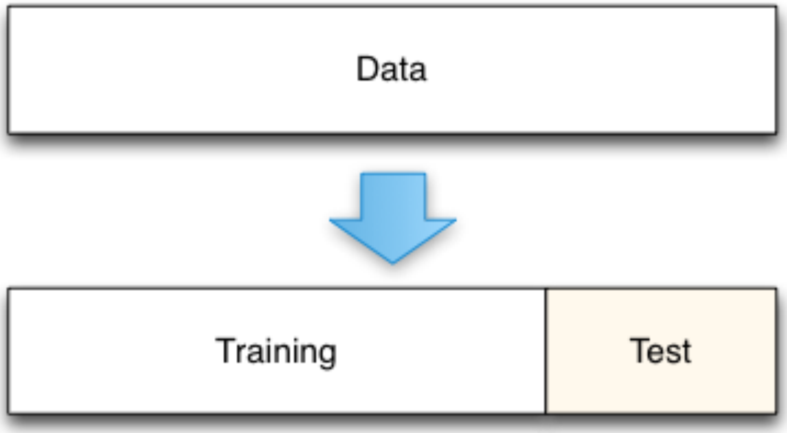
**Word Cloud of Extrovert posts:**
*n=1,999*



# Building the Algorithm

# Training/Testing Set Creation

- Utilized Python's Scikit-Learn machine learning library.
- Dataset was split into striated training and testing sets:
  - Training set = 90% of data
  - Testing set = 10% of data
- Strings were transformed to numeric vectors: CountVectorizer & TfidfVectorizer



# Results

- Each algorithm was trained using the training set.
- Each algorithm was fitted to the testing set to assess model performance.
- Each algorithm was fine-tuned to improve performance.

| Algorithm | Best Test Score |
|---|---|
| Multinomial Naïve Bayes | 77.76%<br>(Count Vectorizer) |
| Logistic Regression | **78.11%***<br>(Tfidf Vectorizer) |
| Support Vector Machines (SVM) | 77.65%<br>(Count/Tfidf Vectorizers) |
| Decision Tree | 77.30%<br>(Count Vectorizer) |
| Random Forest | 76.96%<br>(Count Vectorizer) |

***Best Score**

# Discussion/Recommendations

# Discussion

- A score of 78.11% is not ideal for real-life applications.
- Dataset is highly skewed, too many introverted MBTIs.
- With limited data we're unable to generate a meaningful model:
  - The lack of a unique user identifier provides no knowledge potential duplicates.
  - No socio-economic, demographic or time-series data provided.

# Recommendations

- Review user posts with higher scrutiny and remove remaining stop words.
- Consolidate words with same root word (i.e. “testing”, “tested”, “test”).
- Re-run study with equal amounts of introverted and extroverted participants.
- Collect additional data: demographic information and time spent generating posts.

# Questions?