CAPSTONE 2: PERSONALITY PREDICTION

Vy Nguyen

MBTI Personality System

- The Myers-Briggs -Type- Indicator (MBTI) personality type system that divides everyone into 16 distinct personalities based on their answers to the following questions:
  - Are you outwardly or inwardly focused ?
  - How do you prefer to take in information?
  - How do you prefer to make decisions?
  - How do you prefer to live your outer life?
- For each question, there are two options to choose from. The options are: Introversion (**I**) or Extraversion (**E**), Sensing (**S**) or Intuition (**N**), Thinking (**T**) or Feeling (**F**) and Judging (**J**) or Perception (**P**) respectively. The combination of these options create 16 personality types.
- Challenges: how to know one personality based on their posts on social media?

Goal and Data

- **Goal:** Using Machine Learning models to to predict the personality of a person from the type of posts they put on social media.

- **Data**: The data was collected through the PersonalityCare forum and is available on Kaggle Personality Prediction Dataset.

Data Exploratory

- This dataset consists of over 8675 rows representing 8675 different people and 2 columns representing a person's MBTI personality type and the things they have posted.

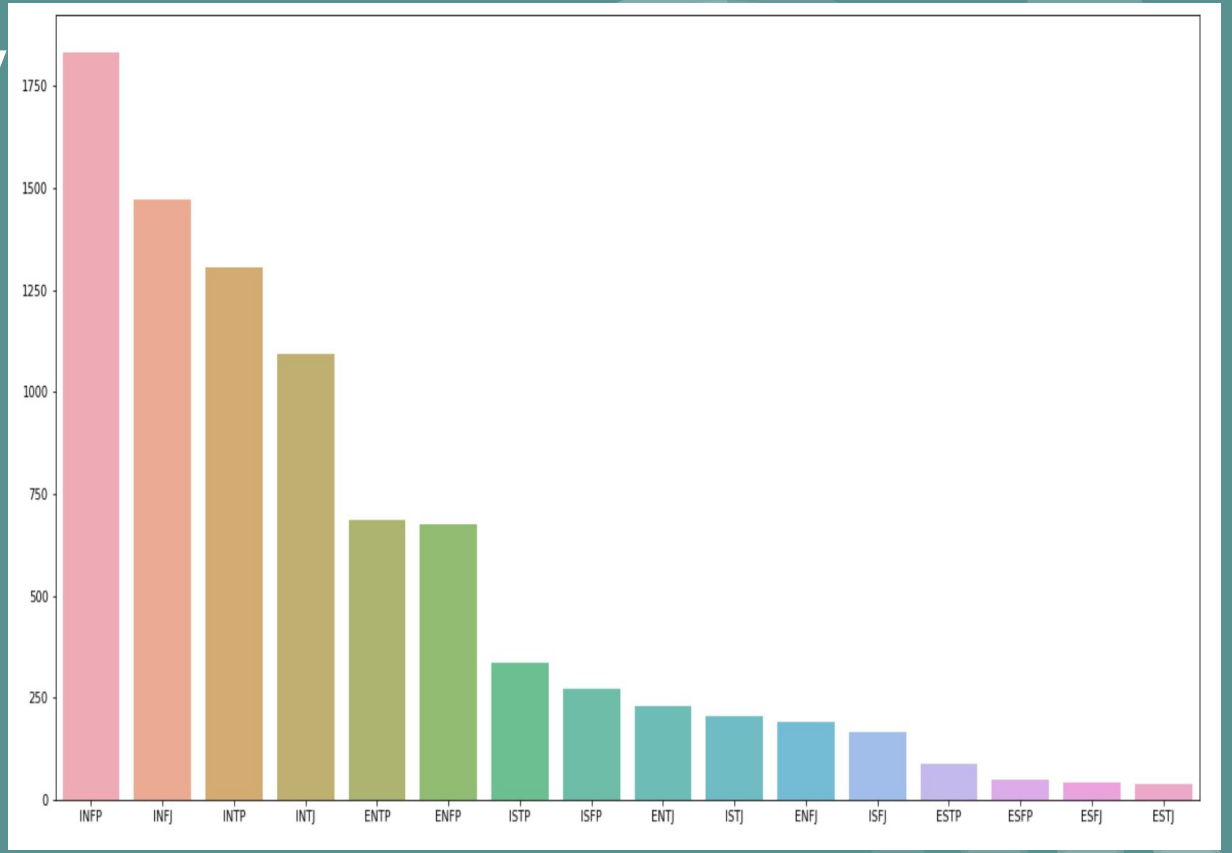
- The target variable is **type** which is the MBTI personality type.

.

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

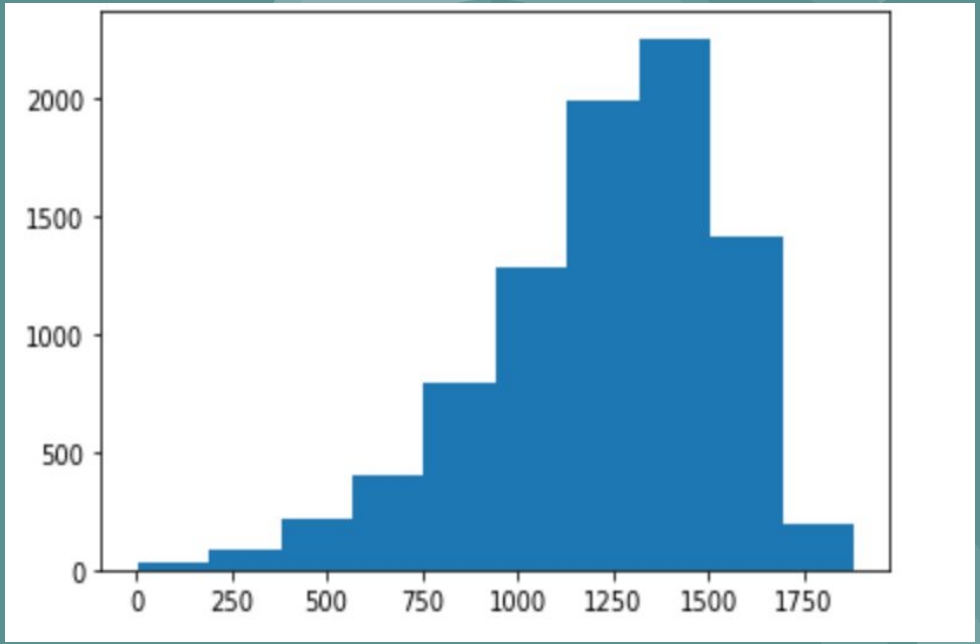
Data Exploratory

- The distribution of personality types is unbalanced.
- The **INFP** with 1832 posts has the highest frequency.
- The **ESTJ** with only 39 posts has the lowest frequency.



Data Exploratory

- the distribution of the length of posts is skewed right.
- The majority posts in our data are longer than 500 words.
- Models based on this data will work best for posts with more than 500 words.



Data Preprocessing

Three functions play a important role in preprocessing text:

- **clean_text** function takes the argument text, the make the text lowercase, remove text in square brackets, remove links, remove punctuation and remove words containing numbers.
- **text_preprocessing** function takes text as an argument. It first applies the function clean_text on the argument, tokenize the result, remove all stop words before combine all tokens.
- **combine_text** function combine single words into text.

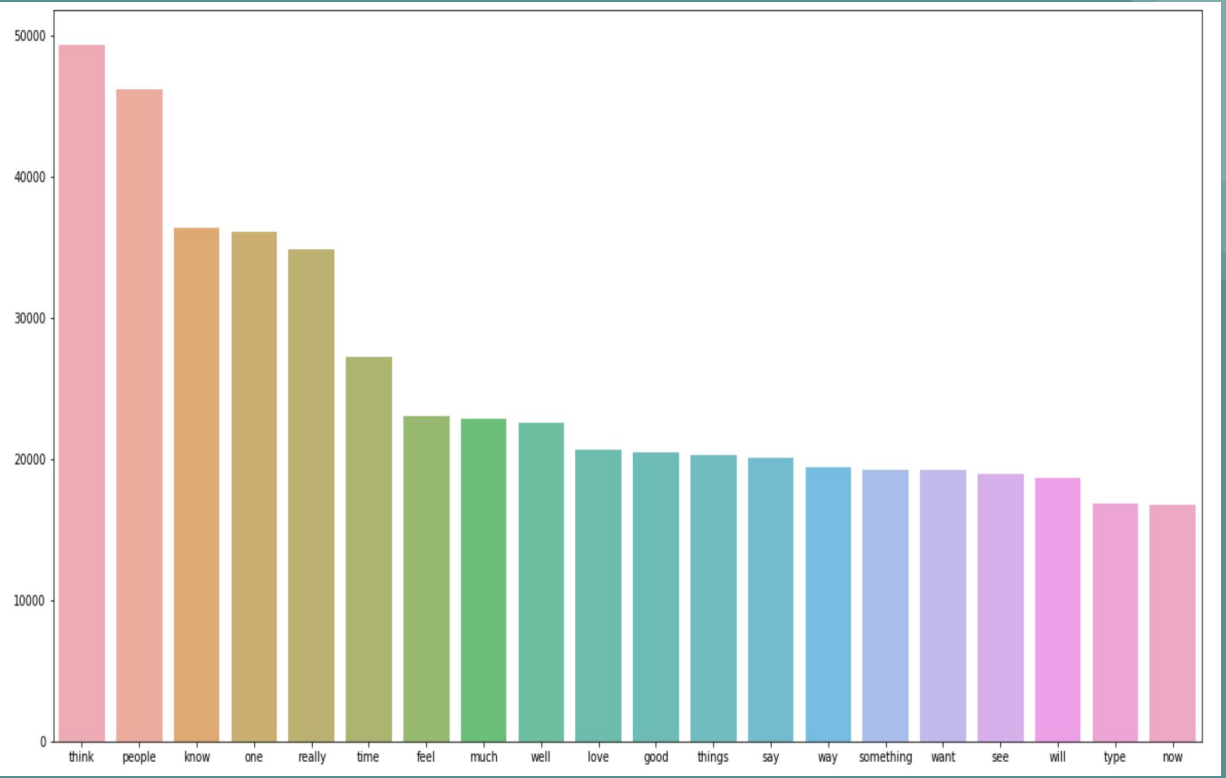
Data Preprocessing

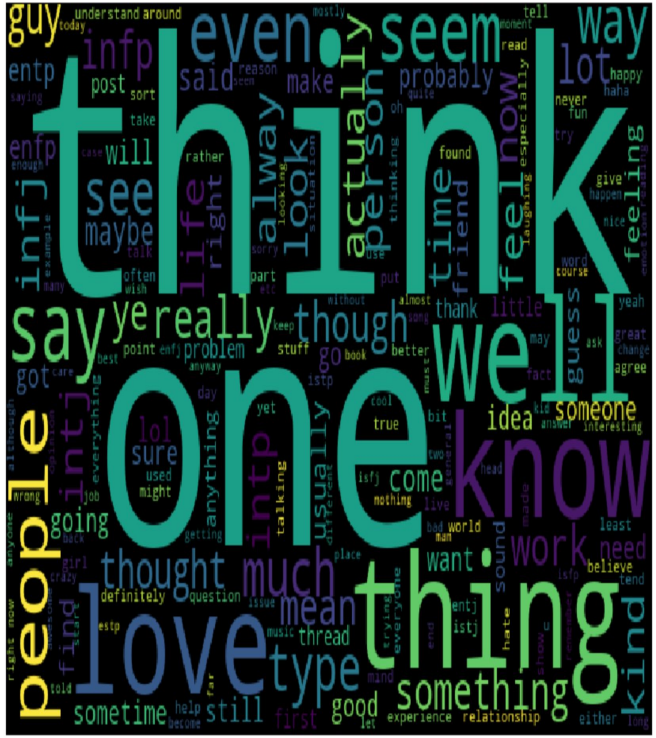
```
0    'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1    'I'm finding the lack of me in these posts ver...
2    'Good one  _____   https://www.youtube.com/wat...
3    'Dear INTP,   I enjoyed our conversation the o...
4    'You're fired.|||That's another silly misconce...
5    '18/37 @.@|||Science  is not perfect. No scien...
```

```
0    ' intj moments sportscenter top ten plays pran...
1    'i'm finding lack posts alarming sex boring po...
2    'good one _____ course say know blessing curse...
3    'dear intp enjoyed conversation day esoteric g...
4    'you're fired another silly misconception appr...
5    ' science perfect scientist claims scientific ...
```

Most Used Words and Word Cloud

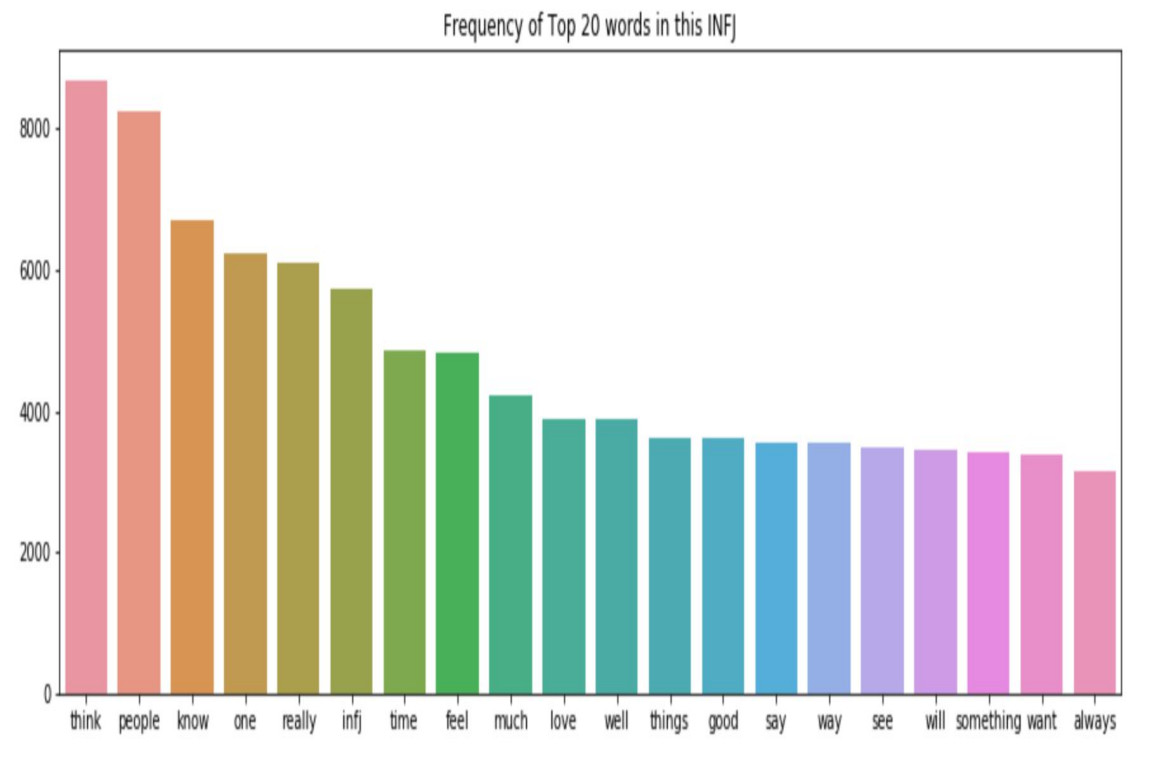
For all posts

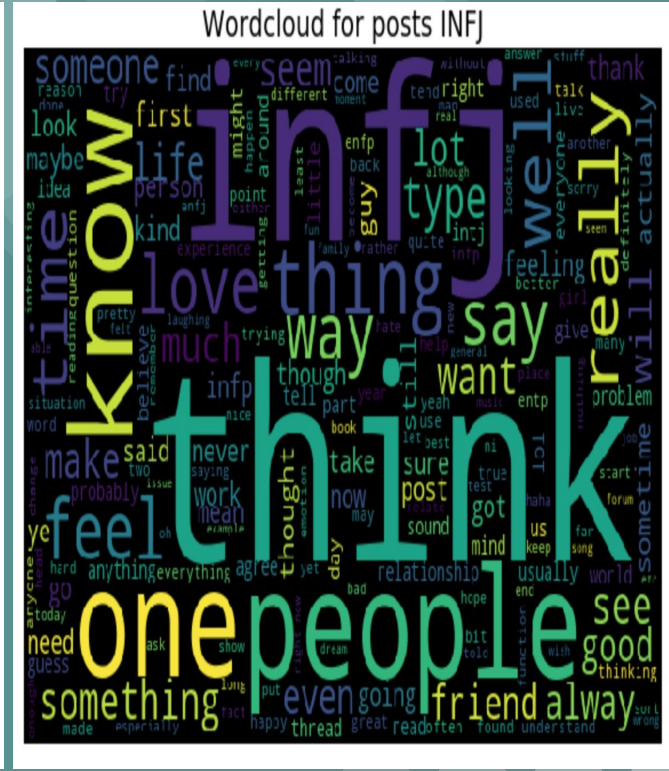




Most Used Words and Word Cloud

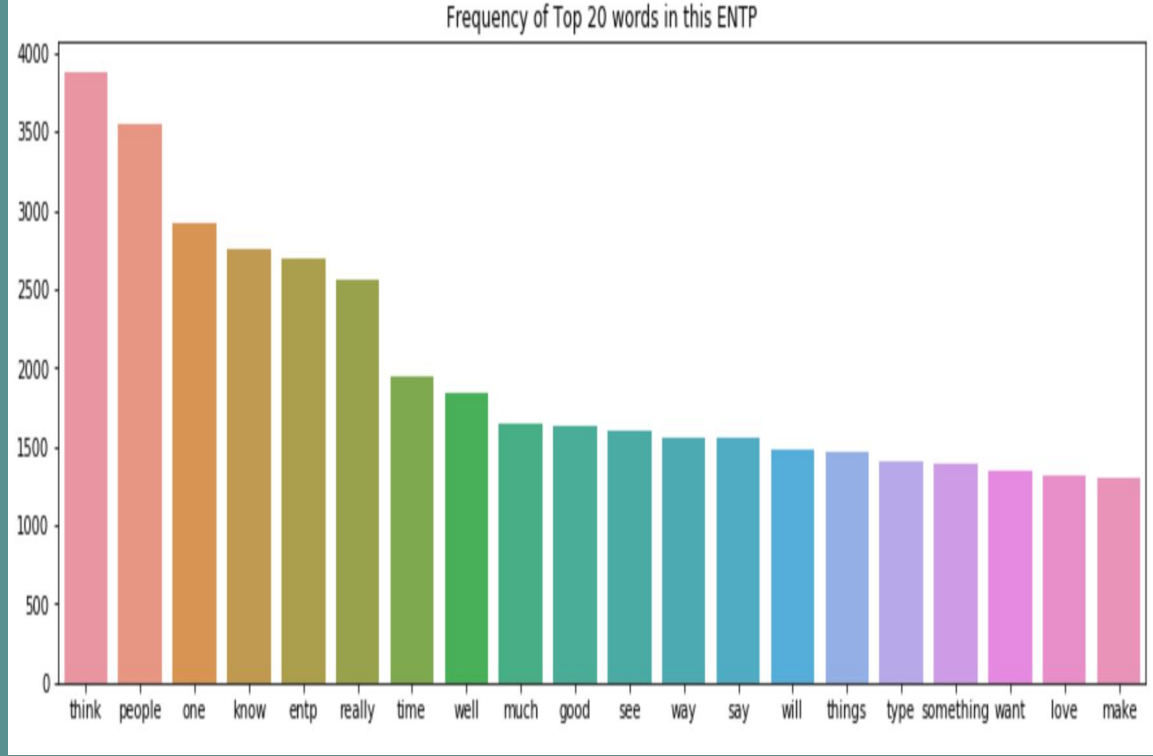
INFJ

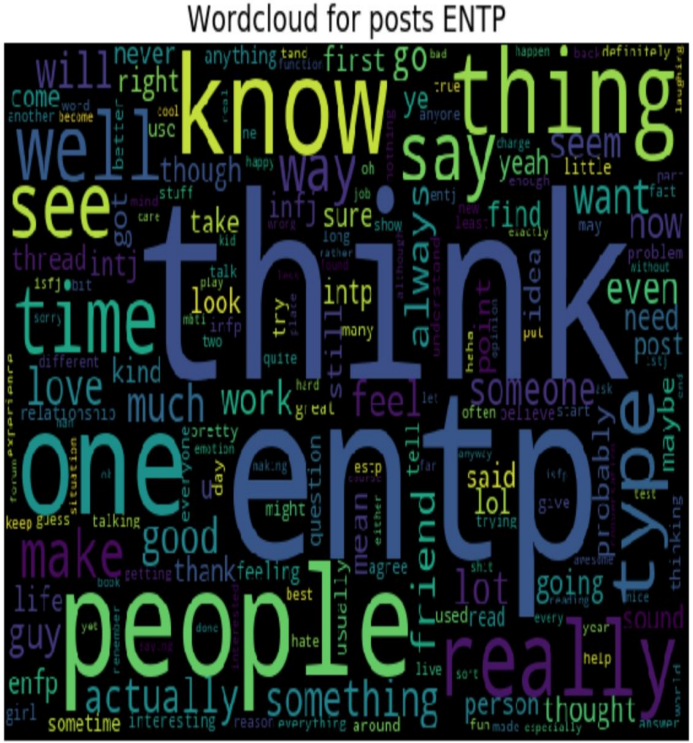




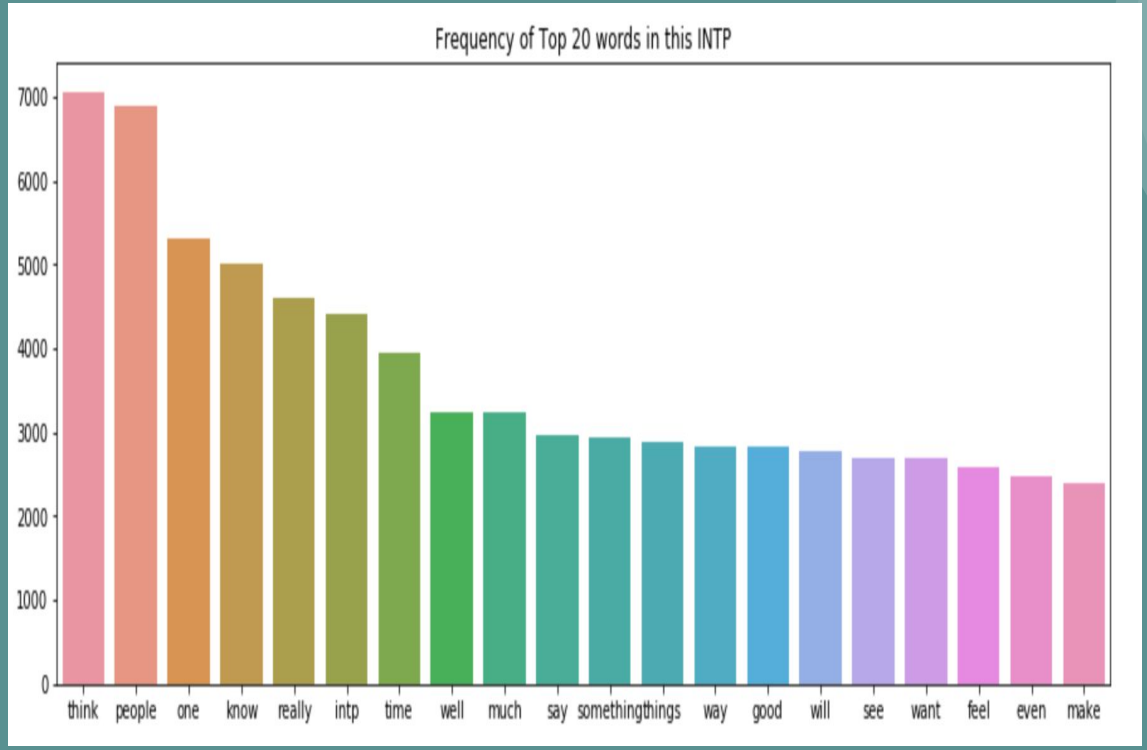
Most Used Words and Word Cloud

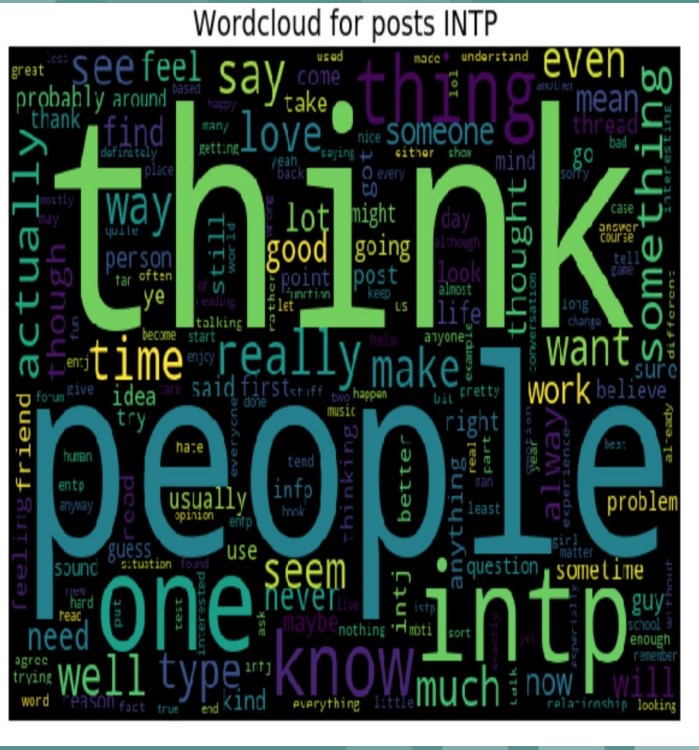
ENTP



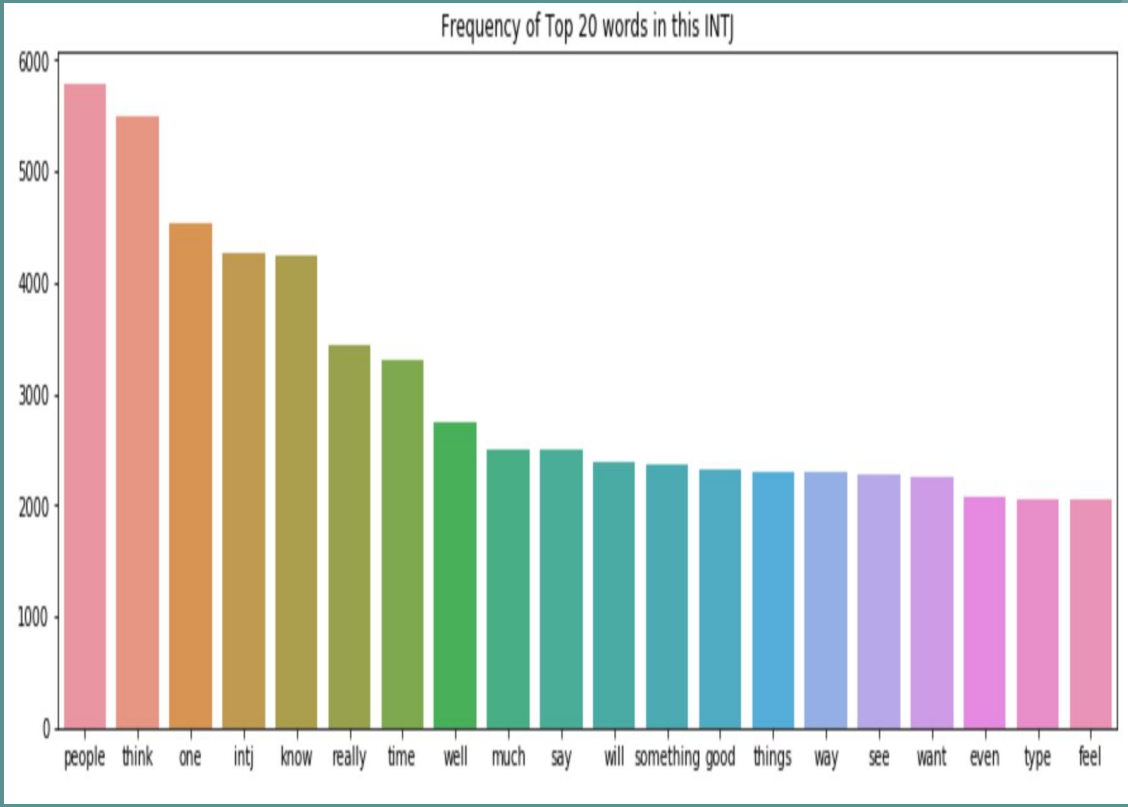


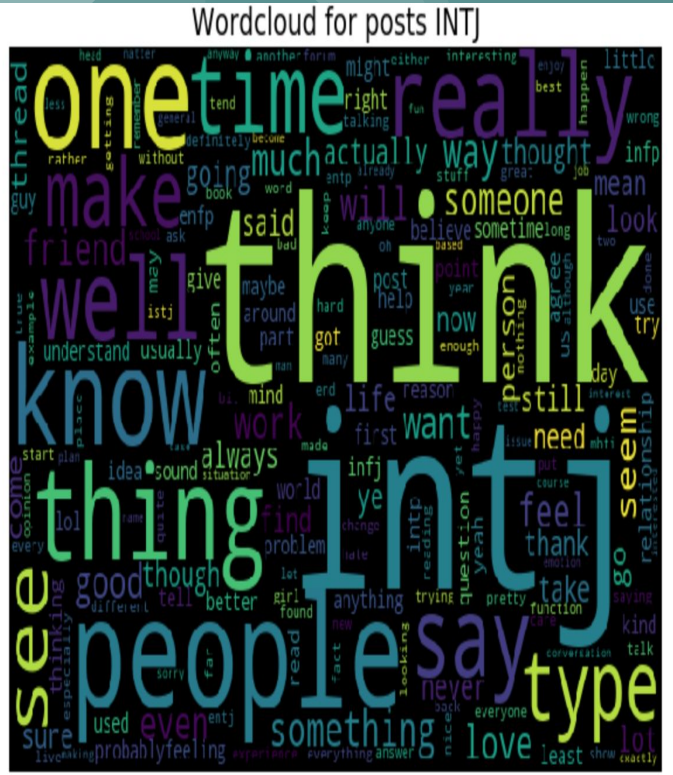
Most Used Words and Word Cloud INTP



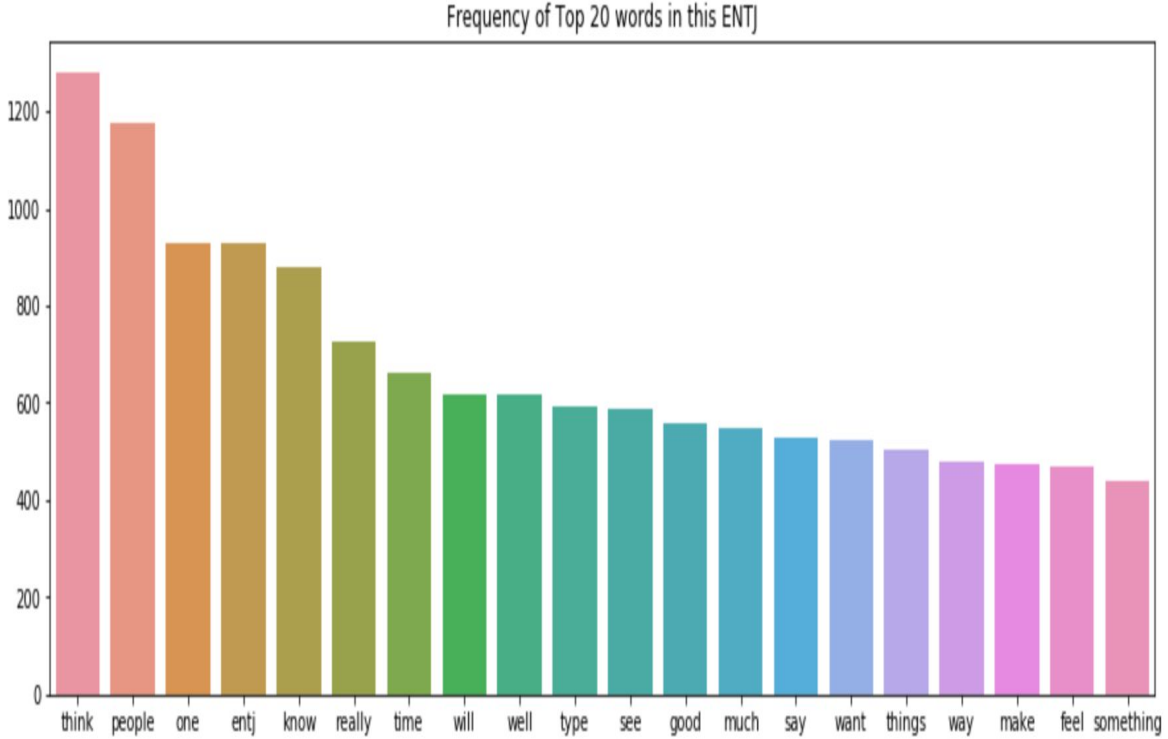


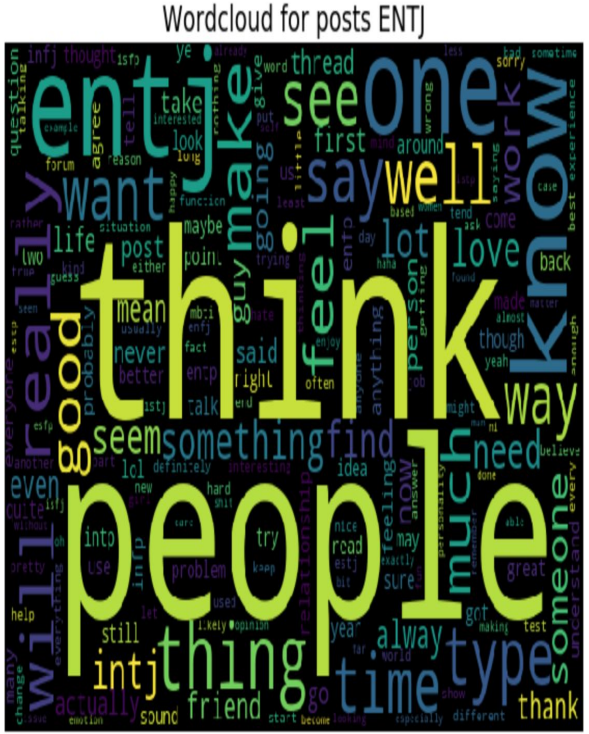
Most Used Words and Word Cloud INTJ



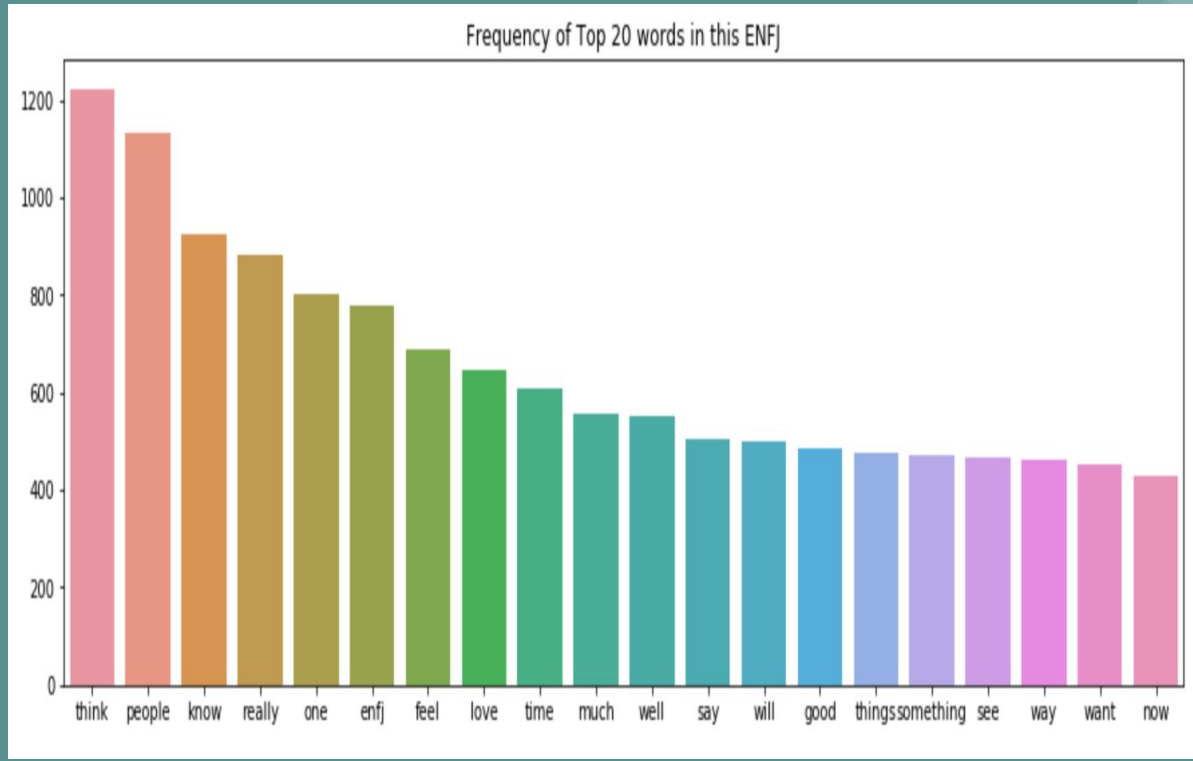


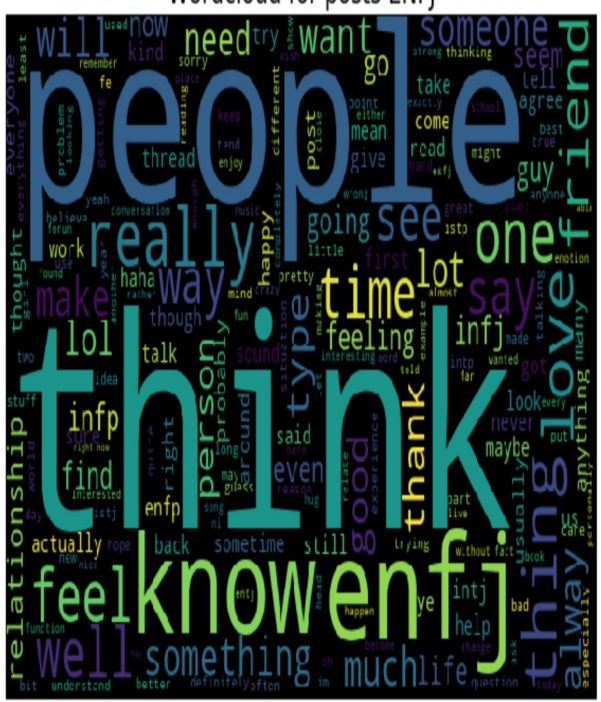
Most Used Words and Word Cloud ENTJ



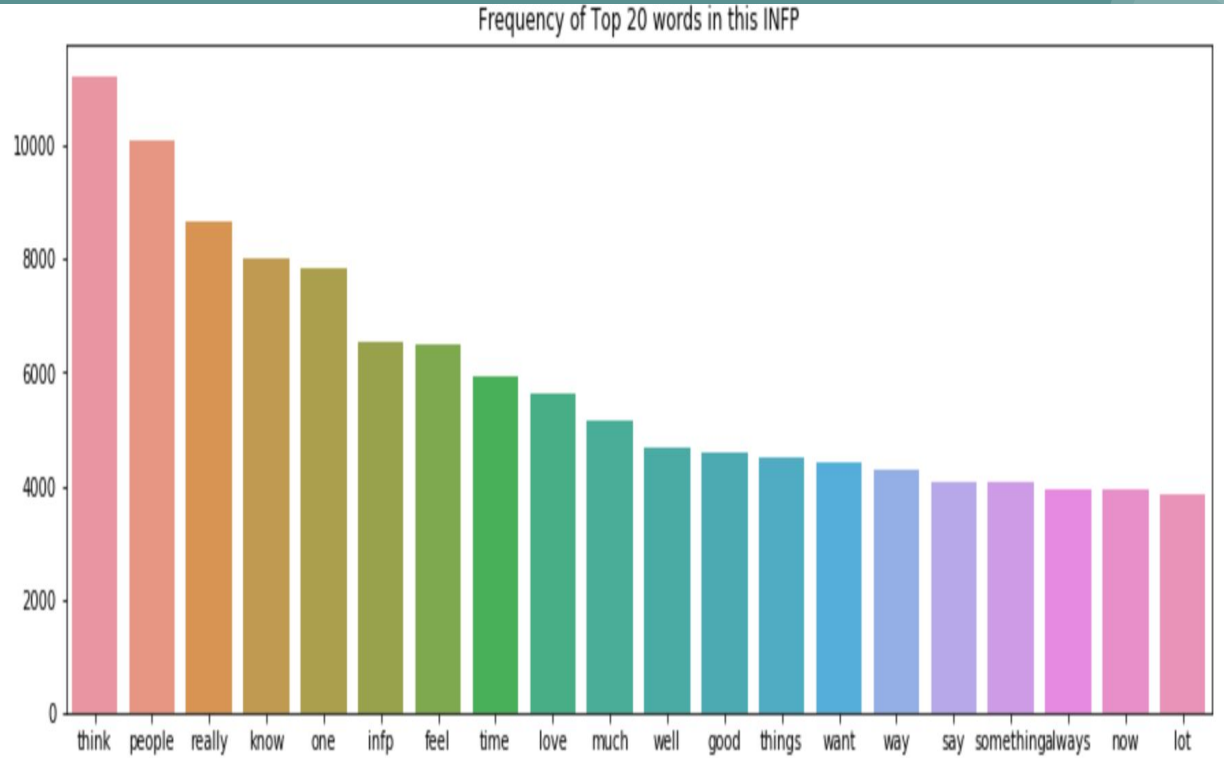


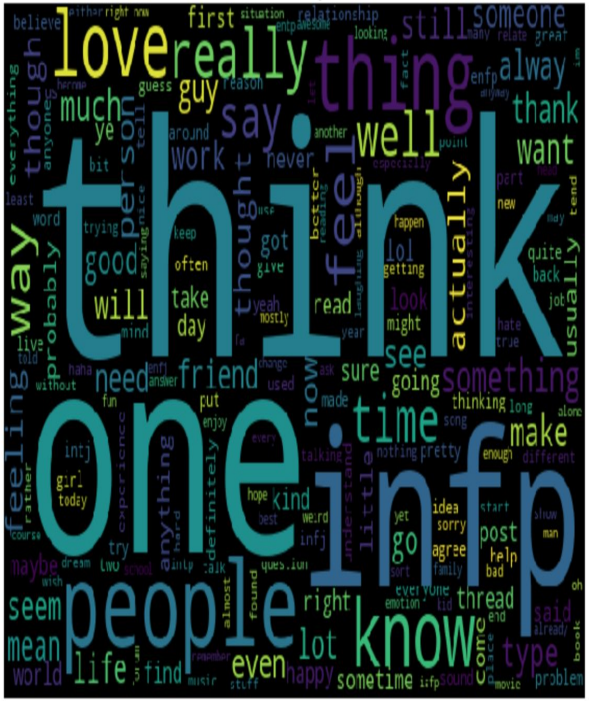
Most Used Words and Word Cloud ENFJ



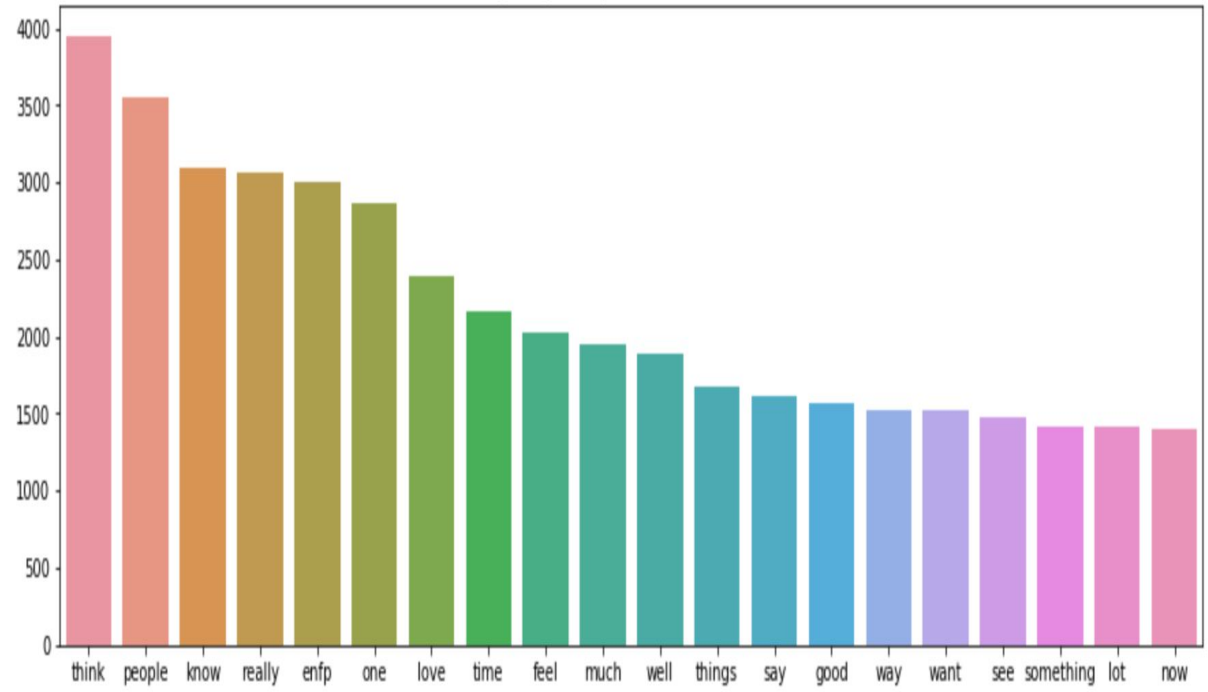


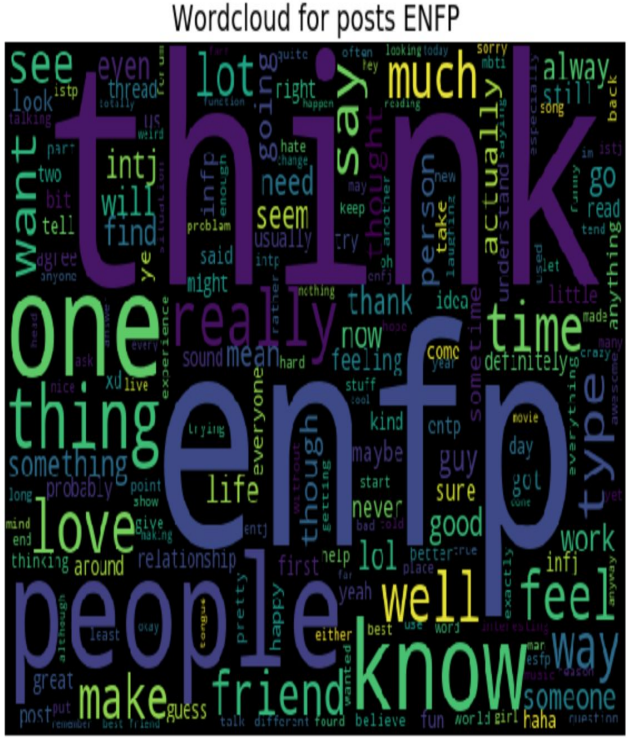
Most Used Words and Word Cloud INFP



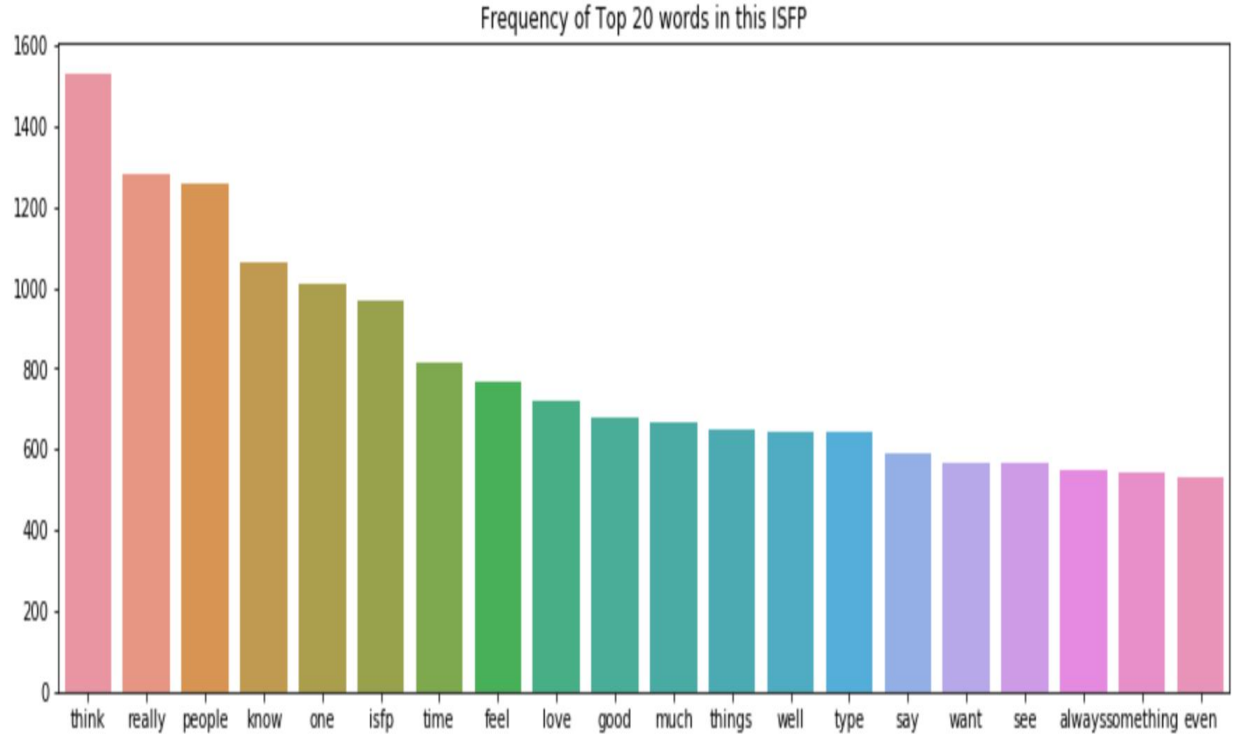


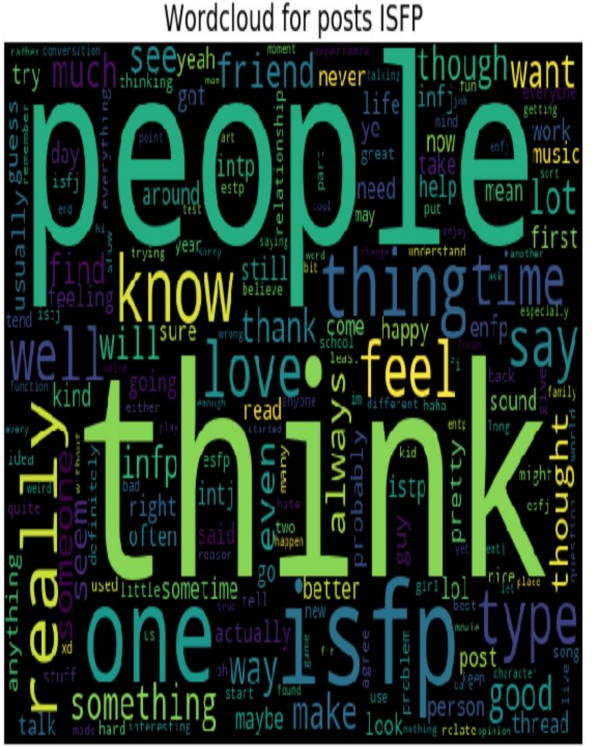
Most Used Words and Word Cloud ENFP



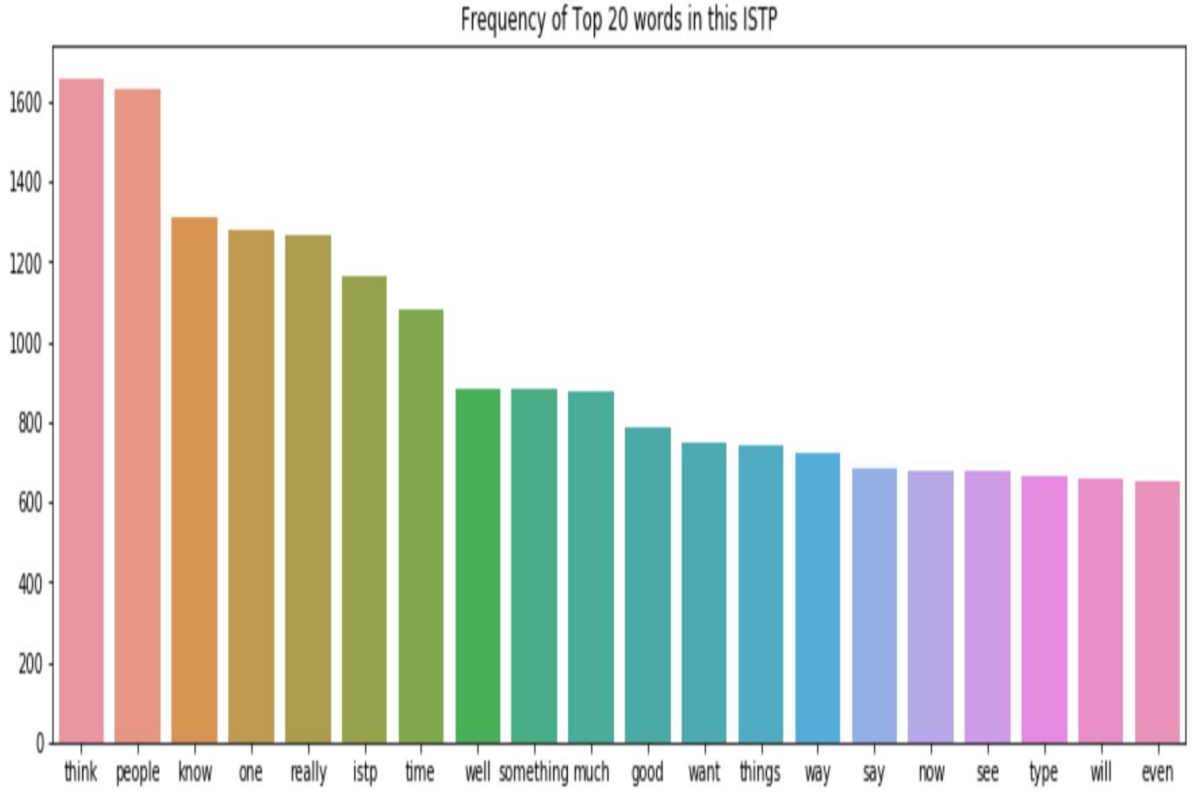


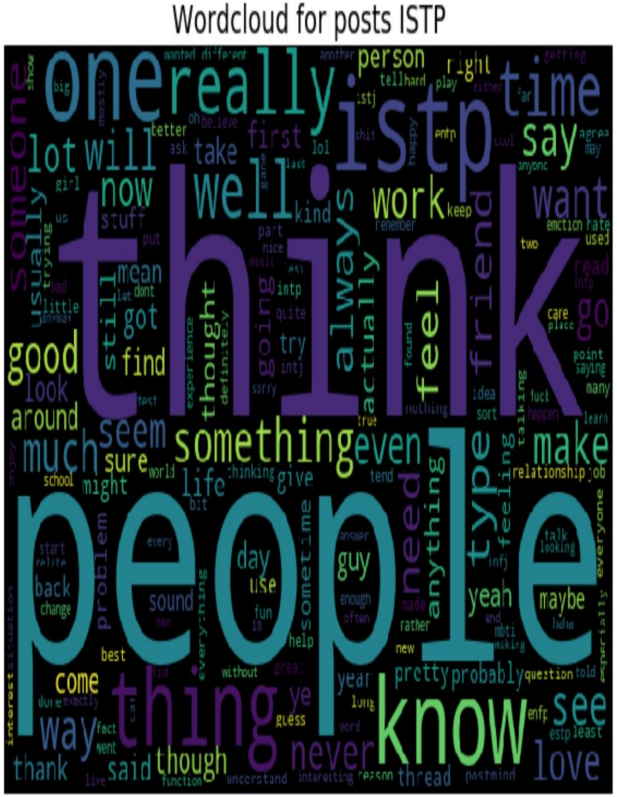
Most Used Words and Word Cloud ISFP



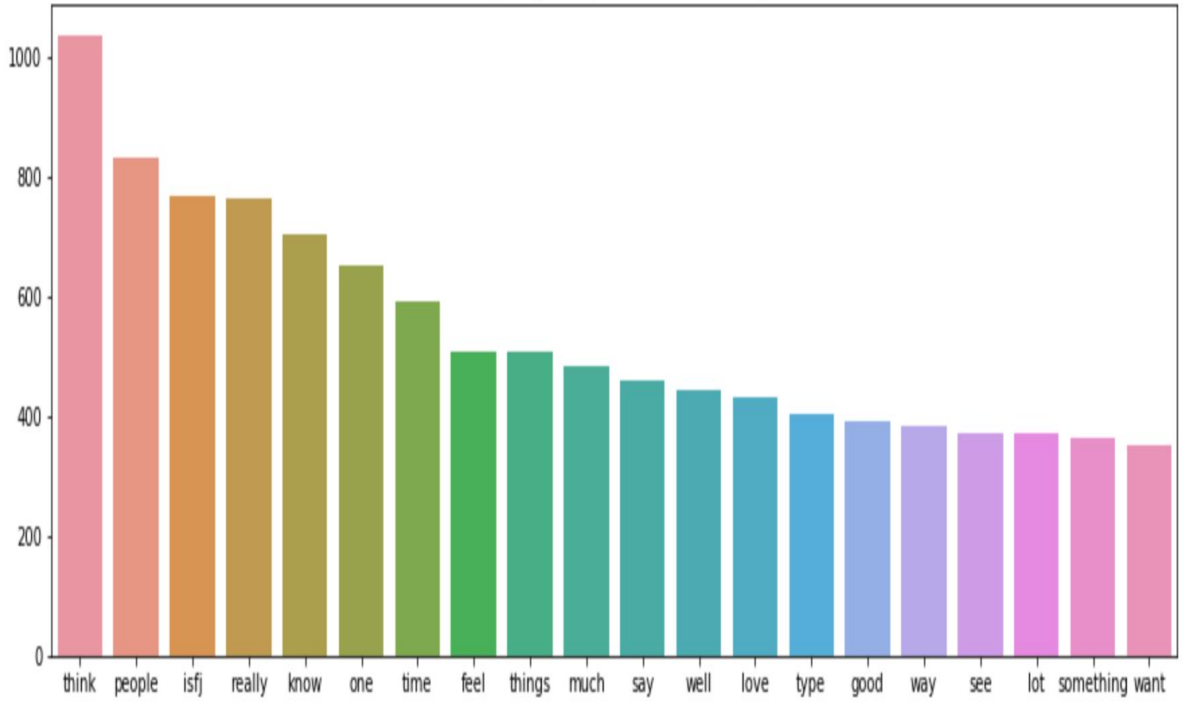


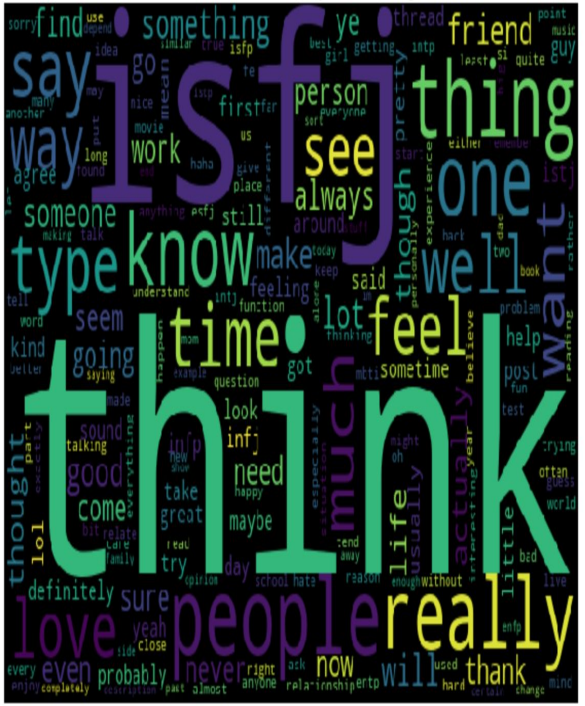
Most Used Words and Word Cloud ISTP



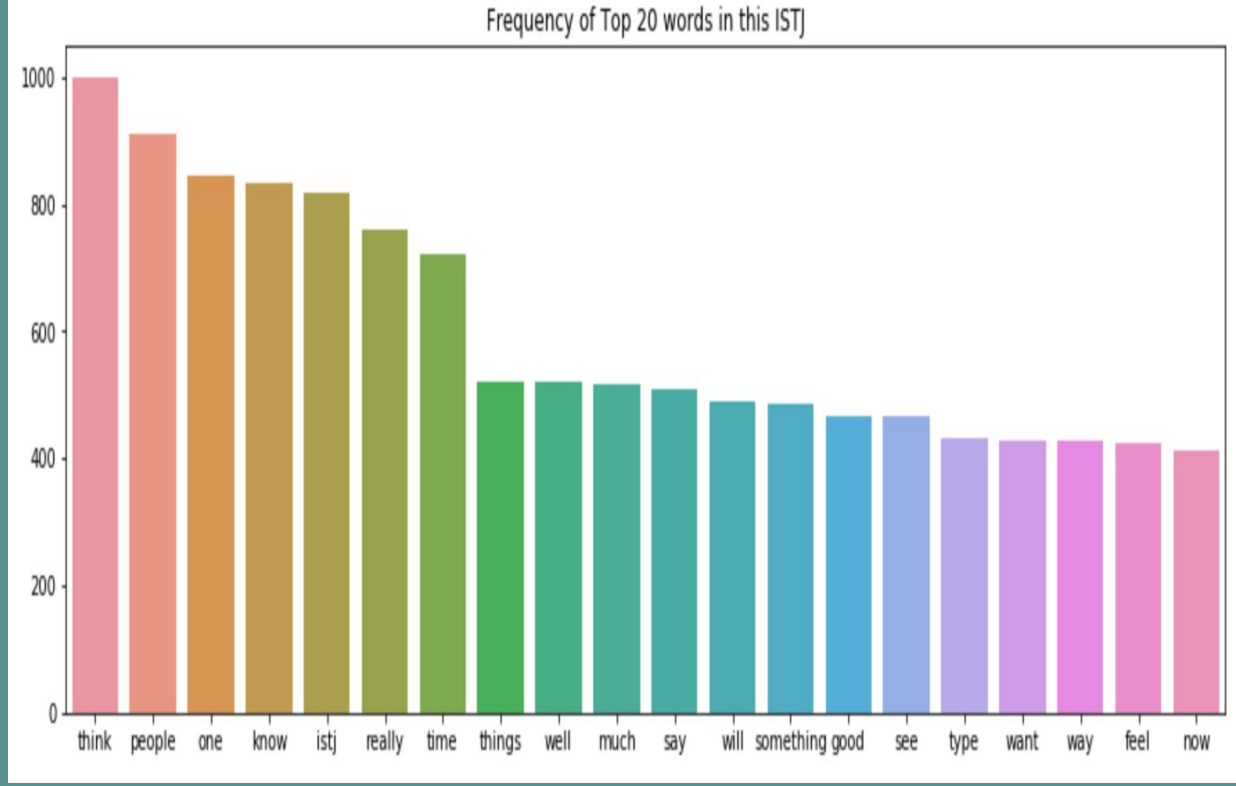


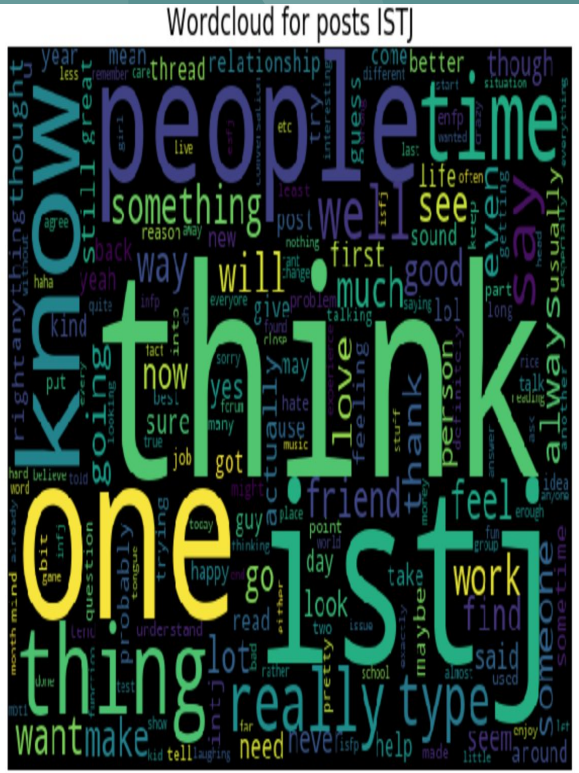
Most Used Words and Word Cloud ISFJ



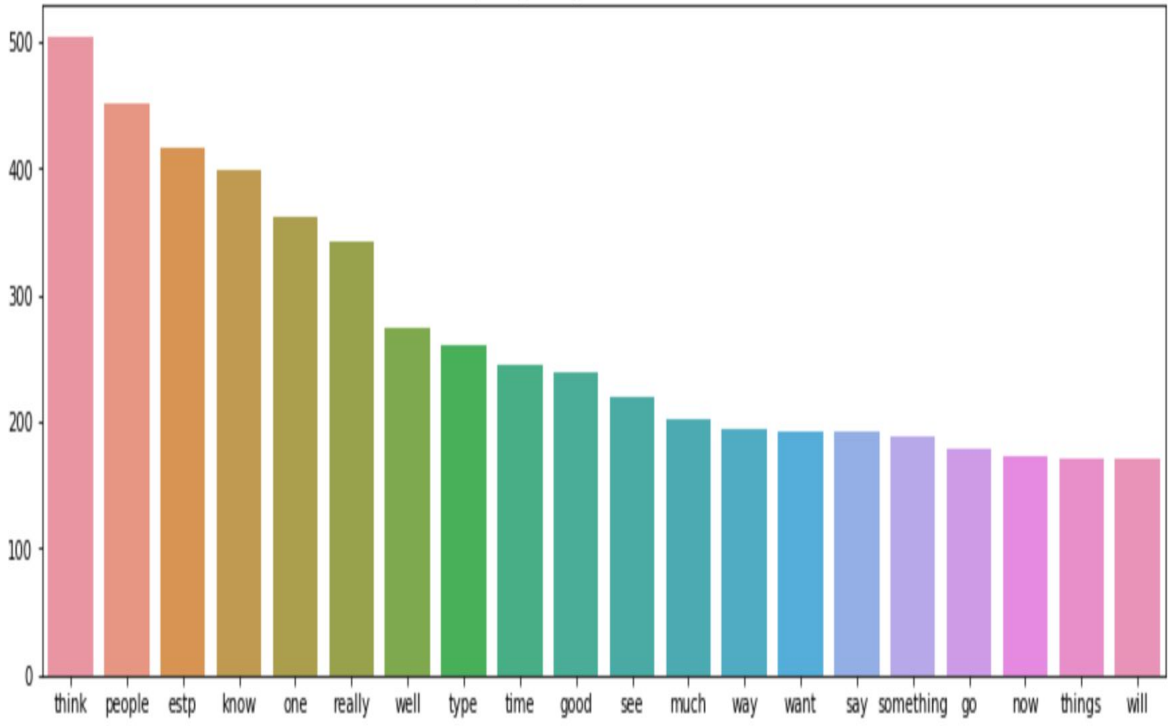


Most Used Words and Word Cloud ISTJ

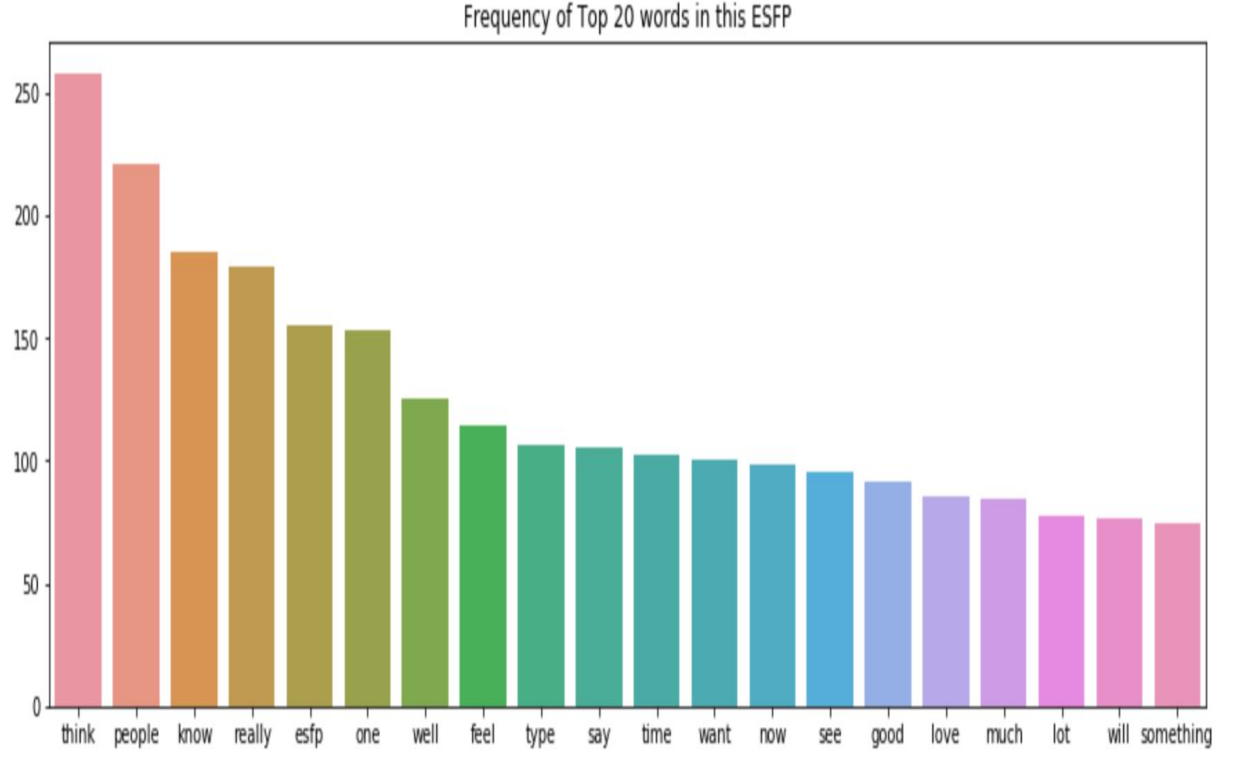


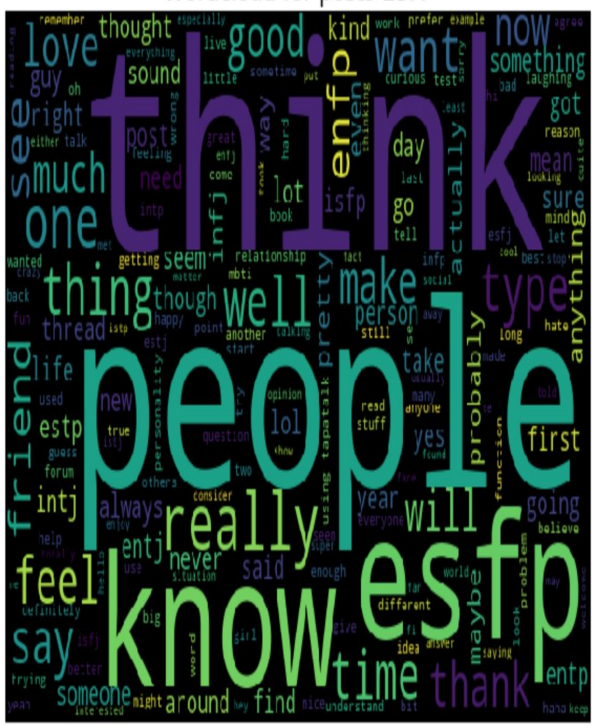


Most Used Words and Word Cloud ESTP

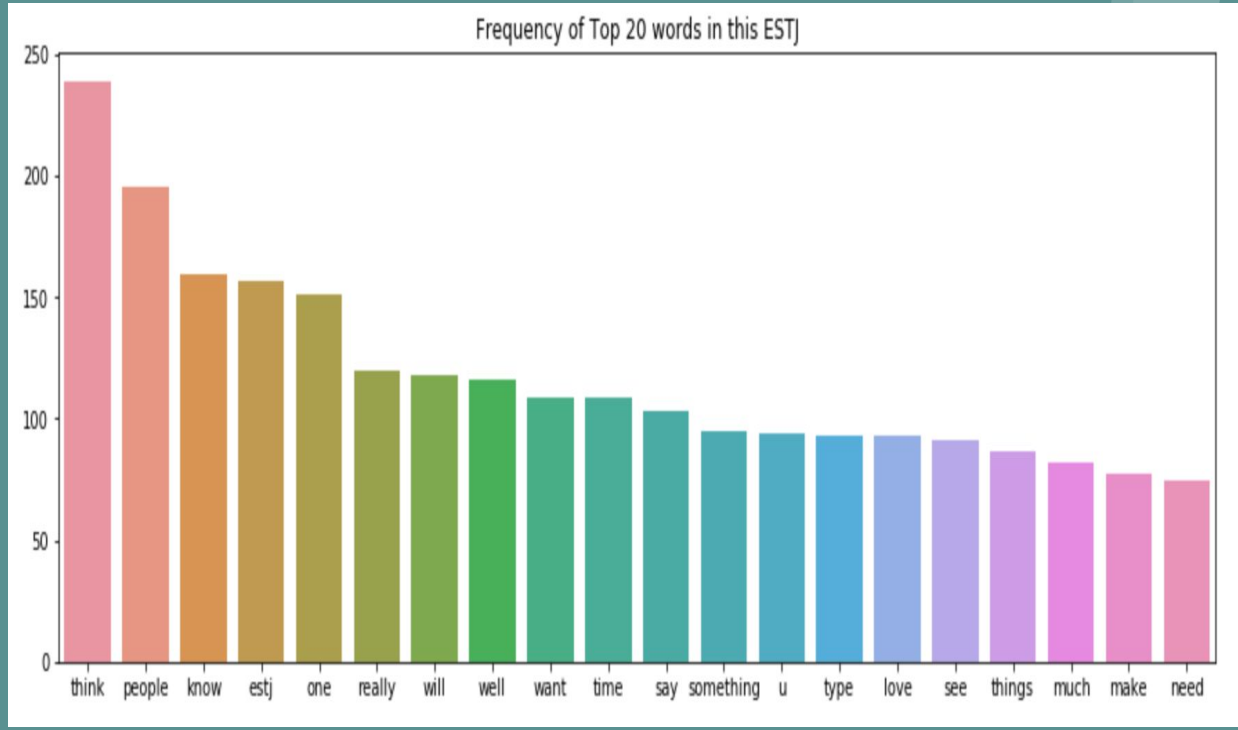


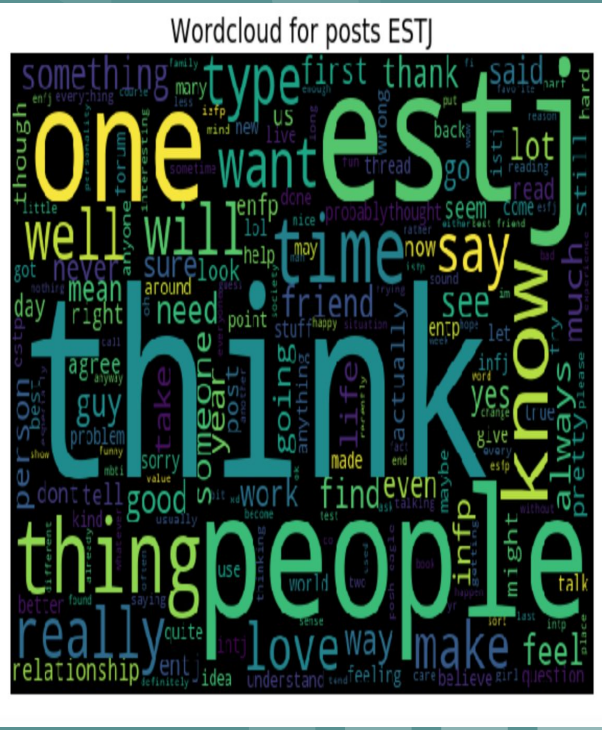
Most Used Words and Word Cloud ESFP



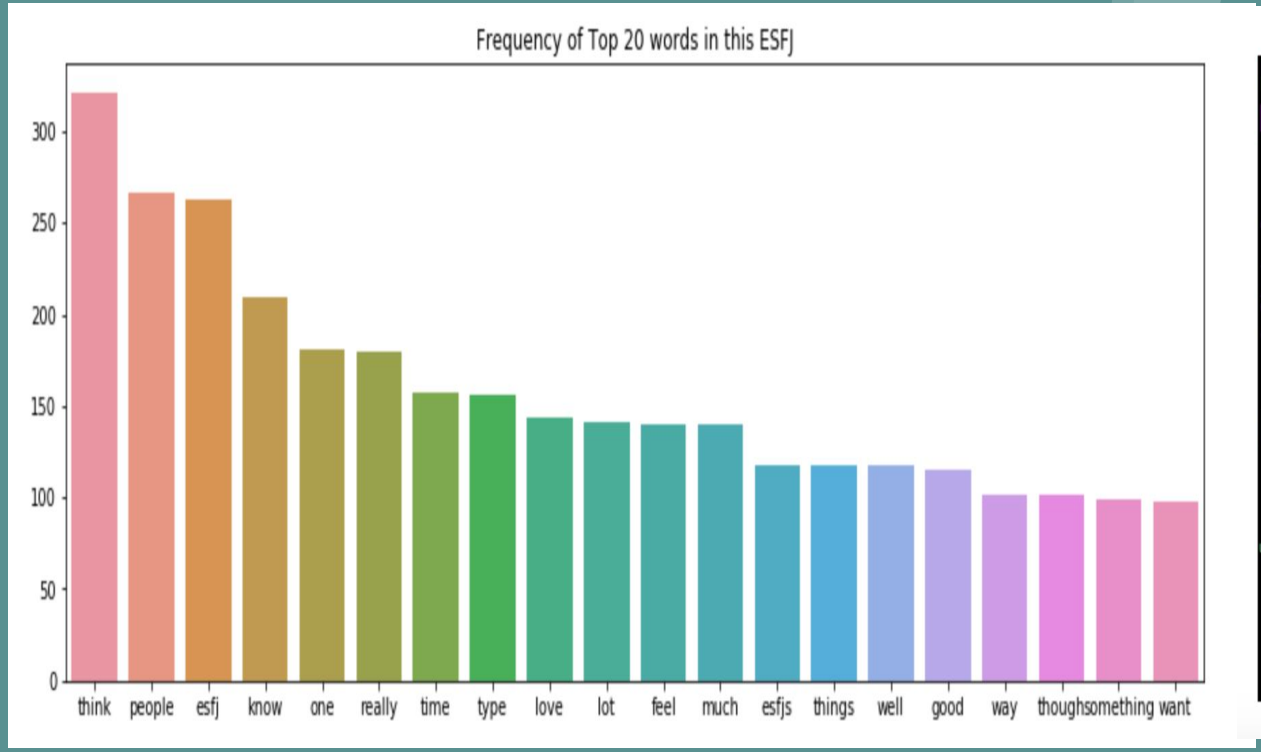


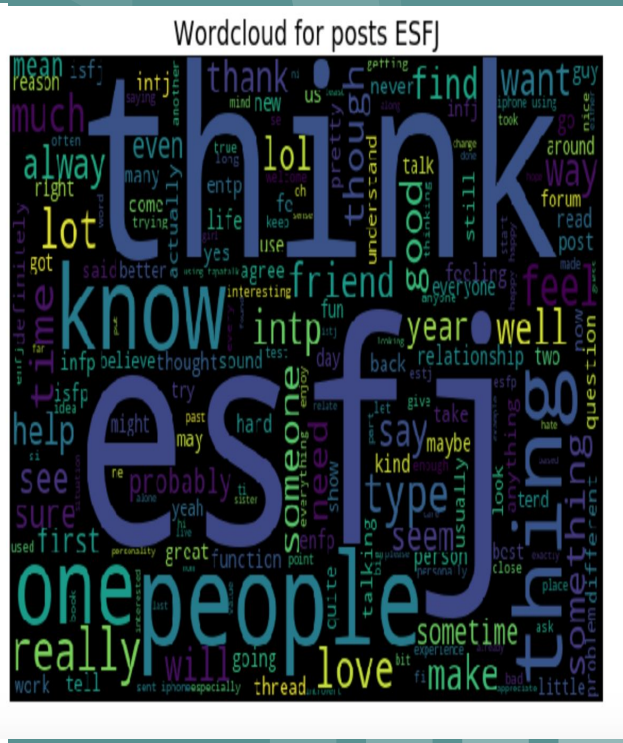
Most Used Words and Word Cloud ESTJ





Most Used Words and Word Cloud ESFJ





Machine Learning

- X is a sparse matrix whose columns are selected words and rows are posts. Each entry in the matrix is the frequency that the word (column) appears in the post(row) which we will use to fit a machine learning model.

- Our problem is a classification problem. We are also performing a category of machine learning which is called supervised learning as we are training our model with a labeled dataset.

- We are using the following Machine Learning Models: K-Nearest Neighborhood, Logistic Regression, Support Vector Machine, Stochastic Gradient Descent, Random Forest, Decision Tree, Naive Bayes.

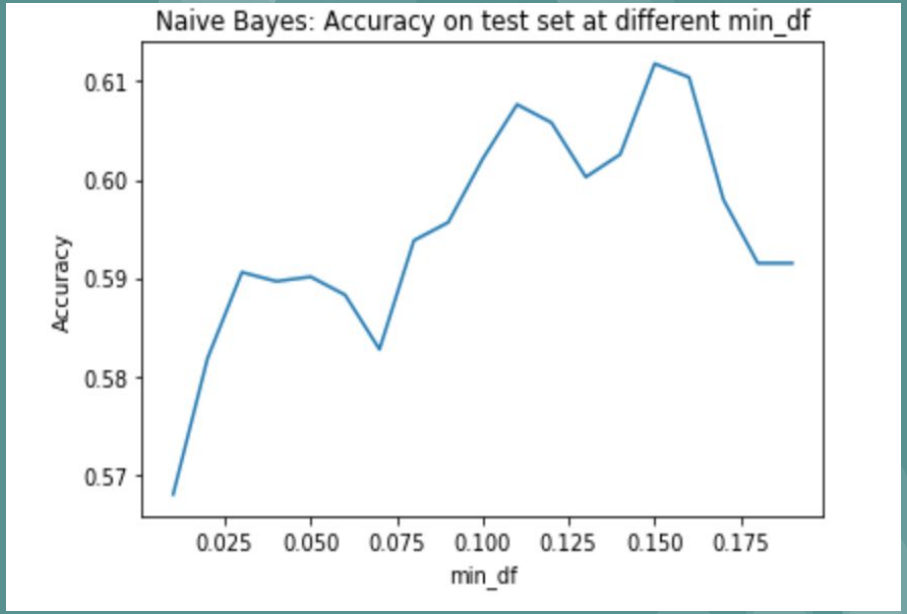
- The metrics to compare these models are: Accuracy, Precision, Recall, F1-score and area under ROC (receiver operating characteristic) curve.

Result

	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.54	0.54	0.54	0.54	0.91
K-Nearest Neighbor	0.36	0.36	0.36	0.35	0.77
SVC	0.56	0.56	0.56	0.56	0.93
Decision Tree	0.47	0.47	0.47	0.47	0.72
Random Forest	0.59	0.60	0.59	0.54	0.91
Stochastic Gradient Descent	0.53	0.56	0.53	0.53	0.80
Naive Bayes	0.60	0.62	0.60	0.61	0.92

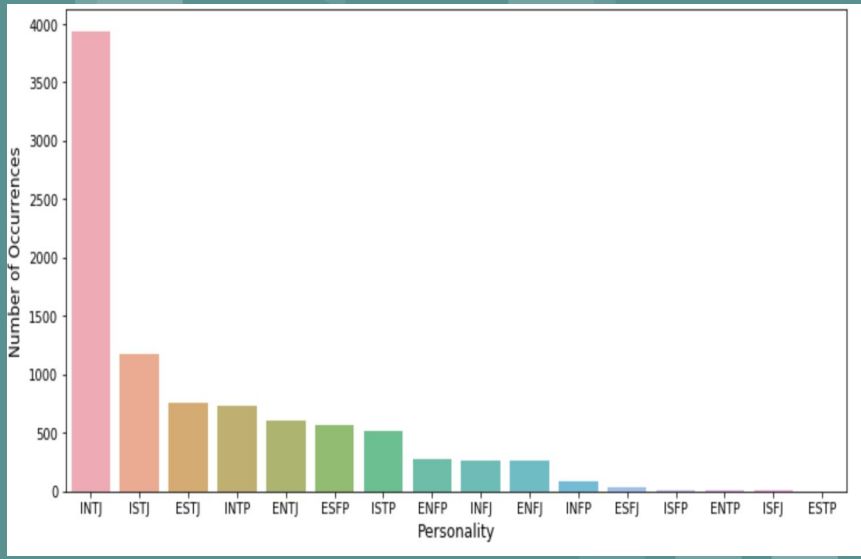
More about Naive Bayes

- We vary the values of min_df and max_df and apply the best model we have(Naive Bayes).
- The highest accuracy is 0.612 when we only consider words that appear in more than 15% and less than 85% of the posts.



Predict Personality of Kaggle Users

- We are going to apply our best model to find out Kaggle users personality based on their comments on Kaggle. The ForumMessages.csv is available on Kaggle at https://www.kaggle.com/kaggle/meta-kaggle?select=ForumMessages.csv
- We observe that the majority of kaggle users have INTJ personality type.



Predict Personality of Public Figures

- There are two US political figures that seem to stand for the opposite point of views: President Donald Trump and previous president Barack Obama. We will collect their posts and use our best model to predict their personality.

	Name	Posts	posts_length	clean_posts
0	Barack Obama	Michelle and I have been spending a lot of tim...	766	michelle spending lot time together past month...
1	Donald Trump	No doubt many people told him his vision wasn...	1146	doubt many people told vision wasn t possible ...

- Based on our best model, they are both predicted to have INFP(Introversion-Intuition-Feeling-Perception) personality type.

Conclusion

- Social media posts become golden data if they are studied correctly.
- My project focuses on using machine learning algorithms in supervised machine learning to predict the personality of a person from the type of posts they put on social media.
- Psychologists can use this project to study personality.
- This project can also help companies who wish to learn more about their customers from their social media posts to provide appropriate services.