Evaluating the MBTI Personality Construct Using Text Data

Prepared by Benjamin Olsen

1. Introduction

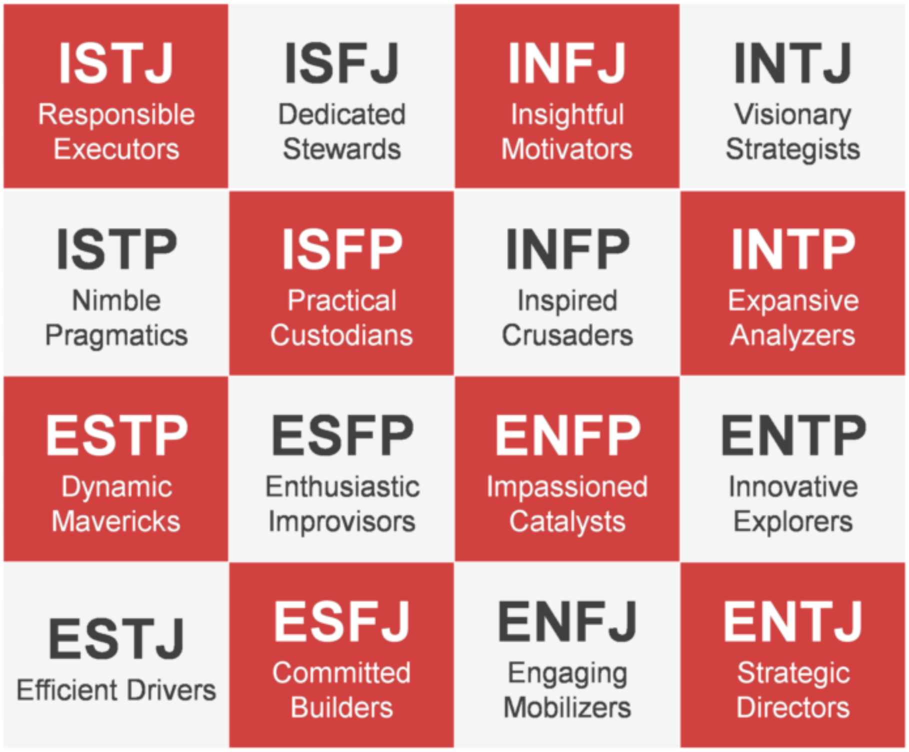
This project is an effort to determine the usefulness of the Myers Briggs Type Indicator as a construct of personality. The Myers Briggs Type Indicator (or MBTI) in short is an "introspective self-report questionnaire with the purpose of indicating differing psychological preferences in how people perceive the world around them and make decisions" according to Wikipedia. While generally accepted by various internet communities as a valid assessment of personality type, the professional field of psychology has discarded this test due to its poor retestibility results. While the clinical application of the test has been discredited as unreliable, the test's usefulness has yet to be evaluated in the context of online behavior. This project will attempt to validate the relevance of the MBTI in an online social context by attempting to classify the text that a person posts online by the MBTI classes. While MBTI is no longer considered to be theoretically validated construction of personality, a distinguishable difference in features of text per reported personality could potentially point to a manifestation of personality in posting style.

2. Background Details

The Myers-Briggs Type Indicator test is a series of questions designed to score individuals along four separate axes. According to their answers, test-takers receive a score for each axis, which represents certain personality aspects. The axis are:

- Introversion (I) – Extroversion (E)
- Intuition (N) – Sensing (S)
- Thinking (T) – Feeling (F)
- Judging (J) – Perceiving (P)

Each axis end is assumedly opposite in characteristic to its constituent on the other end of the axis, and each axis is split down the center. If a person scores to one side of the center, they are assigned the attribute for the side they scored closest to. For instance, on the I (Introvert) to E (Extrovert) scale, someone can score closer to I than to E, and they are assigned the letter ‘I’ for that scale, indicating that they are more introverted than extroverted. Once the scores from all the axis are completed, the test-taker receives a string of letters that represents their personality traits. For example, an INFP is considered to be introverted, intuitive, feeling, and perceiving. For a more in depth description of the classes, please visit <https://en.wikipedia.org/wiki/Myers%E2%80%93Briggs_Type_Indicator>.



ISTJ Responsible Executors	ISFJ Dedicated Stewards	INFJ Insightful Motivators	INTJ Visionary Strategists
ISTP Nimble Pragmatics	ISFP Practical Custodians	INFP Inspired Crusaders	INTP Expansive Analyzers
ESTP Dynamic Mavericks	ESFP Enthusiastic Improvisors	ENFP Impassioned Catalysts	ENTP Innovative Explorers
ESTJ Efficient Drivers	ESFJ Committed Builders	ENFJ Engaging Mobilizers	ENTJ Strategic Directors

Figure 1. All 16 possible combinations of the 4 MBTI axes.

3. Controversy

The Myers-Briggs Type Indicator (MBTI) test faced scrutiny due to poor retest. What was found was that often times, upon taking the test, test-takers would receive slightly different scores. When added to other test-takers results, these scores were normally distributed around the center of the axis. Since the axis is split down the middle, this meant that upon retesting, a slight difference in score could lead to someone being dropped into the opposite class from the last time they took the test, making the test unreliable. This highlighted the dynamic nature of personality and called into question the validity of the MBTI as a personality construct. Modern iterations of personality tests attempt to assign personality attributes by more generative methods, rather than discriminative. Survival of the MBTI personality construct is mostly due to its practical administration methods and easy-to-understand breakdown of the personality types. It lives on in many different iterations online, where it is quite popular.

4. A Secondary Approach

Like mentioned above, the MBTI personality construct is quite popular online, meaning that while other more clinical methods may be much more reliable, the MBTI has the advantage of being more mainstream, and thus also has much more readily available data. The data set in question that we will be analyzing contains over 8600 rows, in each of which is a person's

“Type” (This person’s 4 letter MBTI code/type), and a section of each of the last 50 things they have posted (Each entry separated by "|||" (3 pipe characters))

Utilizing the raw text data, this analysis attempts to extract distinguishable features of the text that might aid in classifying each of the 16 personality classes. Furthermore, it will attempt gain insight into the data by fitting the extracted text features to a statistical model that can describe the features of the text in relation to the classes and classify them accordingly.

This data was collected through the PersonalityCafe forum, and contains a large selection of people and their MBTI personality type, as well as what they have written. Further information on this data set can be found at <https://www.kaggle.com/datasnaek/mbti-type>.

5. Methods

5.1 Data Wrangling

A disadvantage of working with text data is that there can be quite a few odd and unexpected non-textual elements mixed into the text which make analysis difficult unless these elements are removed. Since the text is collected in the real world and generated by real people online, the text can contain irregular spelling, punctuation, actions denoted by asterisks, emojis, and other irregular elements that can cause problems. To prepare the text data for analysis, these elements must be preprocessed, for which there are a few useful libraries such as `pandas`, `matplotlib`, `nltk`, and `re`. For this particular data set, each document was a collection of 50 comments collected from one user, with each comment separated by |||, so for my first approach, each of the 50 comments in each entry was split into a separate document with the same label, increasing the overall size of the dataset while reducing the number of features per document. I later returned to this step to create another set of the data in which the comments were not separated into individual documents, but instead kept together in one long document, effectively maintaining the original size of the data set with full features (I will touch on this later in the analysis).

The next steps included removing escaping HTML characters that could interfere with code, removing hyperlinks which add no value to the feature extraction, expanding contractions, removing digits, removing all punctuation, and removing stop-words (frequently occuring words with no intrinsic value, such as ‘the’, ‘and’, ‘a’).

After referring to the word clouds of the top words appearing in each axis, I noticed some of the most frequently occuring words were actually the names of the 16 personality classes. (*figures 10 -13*) I returned to this step to create another set of text data which had all mentions of any of the 16 classes by name removed. My reasoning for this step was to simulate text data collected from a different source. Remember that the context for this data set is that the text was collected from a forum dedicated to personality types. What this means is that the data contains an inherent contextual bias, as the people generating the data might not exhibit the same behavior in a space that is not explicitly denoted for discussion of personality types.

For simplicity's sake, I will henceforth refer to the cleaned and split documents as Sample 1, the cleaned and non-split documents as Sample 2, and the cleaned and non-split documents with class names removed as Sample 3. For a more indepth look at the data-wrangling process, please visit: <https://github.com/olsenben/DataScience-Capstone-Project-1/blob/master/MBTI-Dataset-Data-Wrangling.ipynb>.

5.2 Exploratory Data Analysis

Conducting exploratory data analysis (EDA) included checking for skew in representation of the personality classes, checking mean word count for each class, and a word cloud preview of the primary word choices of each class. This was conducted on samples 1 and 2 between all 16 classes, but initial findings were indiscriminate enough to prompt me to reattempt the analysis, investigating based on their axis (two classes for each of four axes) instead of comparing 16 different classes. When doing this however we must understand that we can only compare two ends of the same axis at one time, and not compare results between axis due to feature overlap. The findings were considered by their intrinsic value, assuming that the details hold meaningful information while still attempting to consider the skew in distribution.

5.3 Inferential Statistics

At this step I investigated for differences between mean word count, word count variance, mean number of HTTP links, mean number of question marks, and mean number of exclamation points between each personality type. It was also important to check for any effect of over or under-representation on each of these metrics by calculating the Pearson correlation coefficient between each metric and class entry counts. I revisited each of these metrics again from the perspective of two classes for each of the four axes, although ultimately this step was the least useful in the analysis.

5.4 Feature Extraction

This analysis was conducted using the Bag-of-Words model, which turns each word appearing in the corpus (a cumulative collection of the documents which make up each observation) into a feature by which a machine learning model can be trained. The result is a very sparse feature matrix with *k* features and *n* samples. Each feature is weighted using term frequency–inverse document frequency which normalizes each feature based on a weighting assigned by frequency of appearance. Each document is then converted into a vector, which is essentially a closest fitting line to each feature within a document so that vectors can be compared between classes. This approach is very popular in search engines and also actively being improved upon for sentiment analysis.

5.5 Machine Learning

Initial attempts at fitting the vectorized bag of words to a statistical model included trying Logistic Regression and Multinomial Naive Bayes before settling on a Support Vector Machine trained with Stochastic Gradient Descent, which works particularly well with sparse data-matrices. Looking at the confusion matrix however, it was apparent that there was some overlap between the features of many of the vectorized samples. Returning to the original MBTI construct (4 separate axes), the inherent nature of the classes meant that each sample would share some of its features with at least half of the other classes at any given time. For example, all classes starting with ‘E’ would share overlapping features with one another, adding noise to the model and confusing each class. Also, removing any confounding features would be difficult considering the high number of classes (16 in total).

Further attempts to classify the text features involved training four separate binary classifiers; one for each axis. This was effective for reducing the noise in the data by allowing the classifiers to focus on only two hypothetically polar aspects of the data at a time, and also increased the support for each sample size by decreasing the number of classes it was split into from 16 down to 2 at a time. The issue of the skewed representation of classes in the data was mitigated by balancing the weighting of the classes while fitting the classifier to the training data. Utilizing Scholastic Gradient Descent Classifier with a logarithmic loss function enabled the classifier to predict the probability of a classification given its vectorized text features. Since the probabilities of binary classification are proportional (75% probability of class A infers 25% probability for class B), this approach was effective in simulating the original scoring method in which test scores fall along each axis, but this time the scores represent probabilities of each classification based on vectorized text features classified by a support vector machine.

6. Results

6.1 EDA

From the onset it was quite apparent that there was an extreme skew in the distribution of the data, with certain classes maintaining exponentially higher representation, with the class with the highest entry count at 47x greater than the class with the lowest entry count. This proved to be a constant challenge throughout the rest of the investigation, as it made modeling difficult.

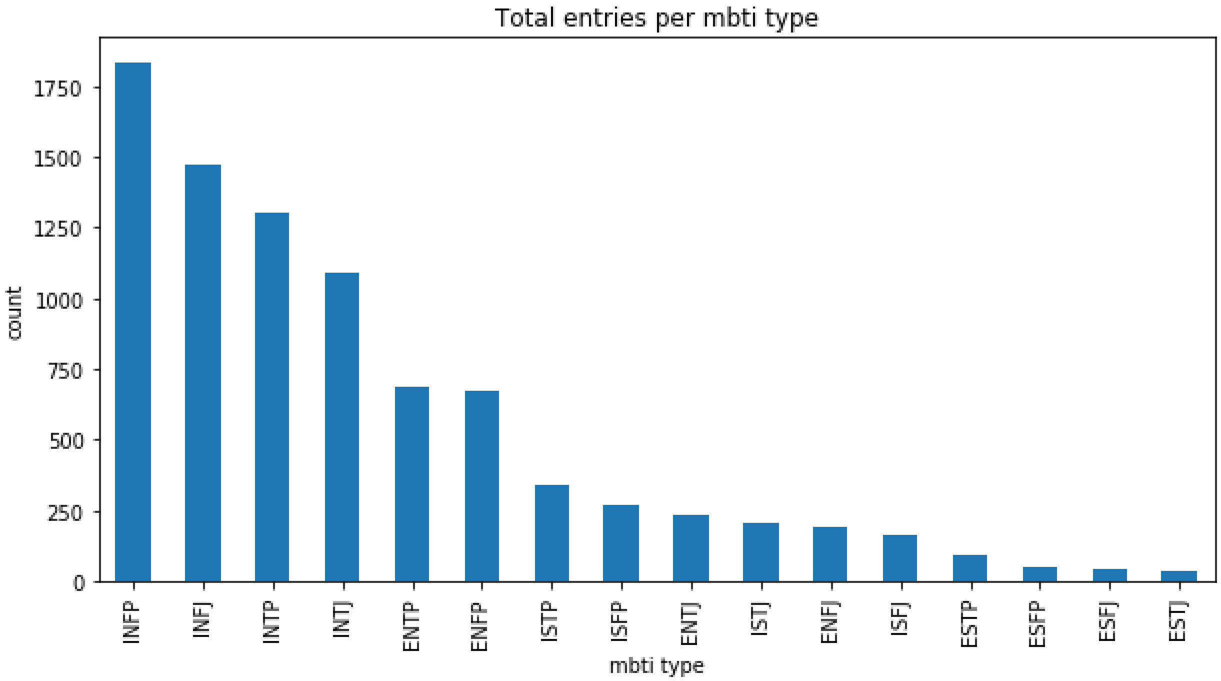


Figure 2. Total entry count of all 16 personality types.

Searching for distinct differences in the mean word count of each entry per personality type found that while there is a difference between each class, it seems to be quite marginal. With 16 classes it is hard to tell (see *figure 14*). Considering how the MBTI tests are classified, at this point it seemed prudent to investigate the data investigate by axis to get a better picture of what is going on. Performing the same checks on the data by axis reduced the amount of comparisons to be made exponentially.

Axis Summary Statistics

Axis	Total entries	Mean word count	std	z-test
Introvert	6675	602.3	142.5	0.405
Extrovert	1999	605.3	136.4	
Intuition	7477	606.5	138.6	8.444e-09
Sensing	1197	581.2	153.9	
Thinking	3981	592.1	140.7	2.700e-11
Feeling	4693	612.3	140.8	
Judging	3434	608.2	141.5	0.006
Perceiving	5240	599.6	140.7	

Figure 3. Summary of EDA statistics by axis.

Interestingly enough, there isn't too much difference within each axis in mean word count. The greatest difference is between intuition and sensing, with a range of 25, but sensing had the lowest entry count out of all the axis classes. The most notable difference in mean word count was between thinking and feeling (592 and 612 respectively), where their entry counts were only different by about 700. Here are some highlights from the analysis:

1. Introverts post 3.33x more often online than extroverts.
2. Those on the Intuition side of the I-S axis tend to post 6.25x more than those on the Sensing side.
3. Those on the Thinking end of the I-F axis post shorter comments by a degree of 3% than those on the Feeling end.
4. Those on the Judging end of the J-P axis tend to post about 1.5x more often than those on the Perceiving side.

For the full analysis and explanation, please visit https://github.com/olsenben/DataScience-Capstone-Project-1/blob/master/MBTI-Data-Storytelling.ipynb

6.2 Inferential Statistics

6.2a 16 classes

My first approach to this data set was purely from the perspective of 16 distinct personality classes. Since Sample 1 wound up not being particularly useful in the final analysis, I will exclude the statistics from Sample 1 and include only the statistics from Sample 2. Sample 3 is also excluded since its only difference from Sample 2 is a lack of class mentions and is only really relevant as a feature in fitting the data to a statistical model.

The mean word count compared between classes was fairly indistinct, like mentioned above *(figure 14*), with the total mean across all 16 classes landing at 595.0 words and a range of 121.8. A Pearson correlation coefficient calculated between total entry count and mean word count returned $r = 0.270$ (*figure 15*). Investigating word count variance instead was helpful to tease out differences between the classes (*figure 16*). To investigate the effects of the skew, the Pearson correlation coefficient between variance and entry count was calculated at $r = -0.40$, showing a mild negative linear relationship (*figure 17*).

The mean number of HTTP links appearing in the text of each class was also hardly distinct and suffered from the effects of a mild positive correlation between entry count ($r = 0.32$) (*figure 18* and *19*).

The same could be said of mean number of question marks, which, while thoroughly unaffected by representation bias from a linear perspective ($r = 0.06$), yielded similarly indistinct differences and no notable grouping trends from a practical standpoint (*figure 20* and *21*).

The mean number of exclamation points between each personality type was the most promising metric with a range of 12 across all 16 classes and only a weak negative correlation with representation bias ($r = -0.23$) (*figure 22* and *23*).

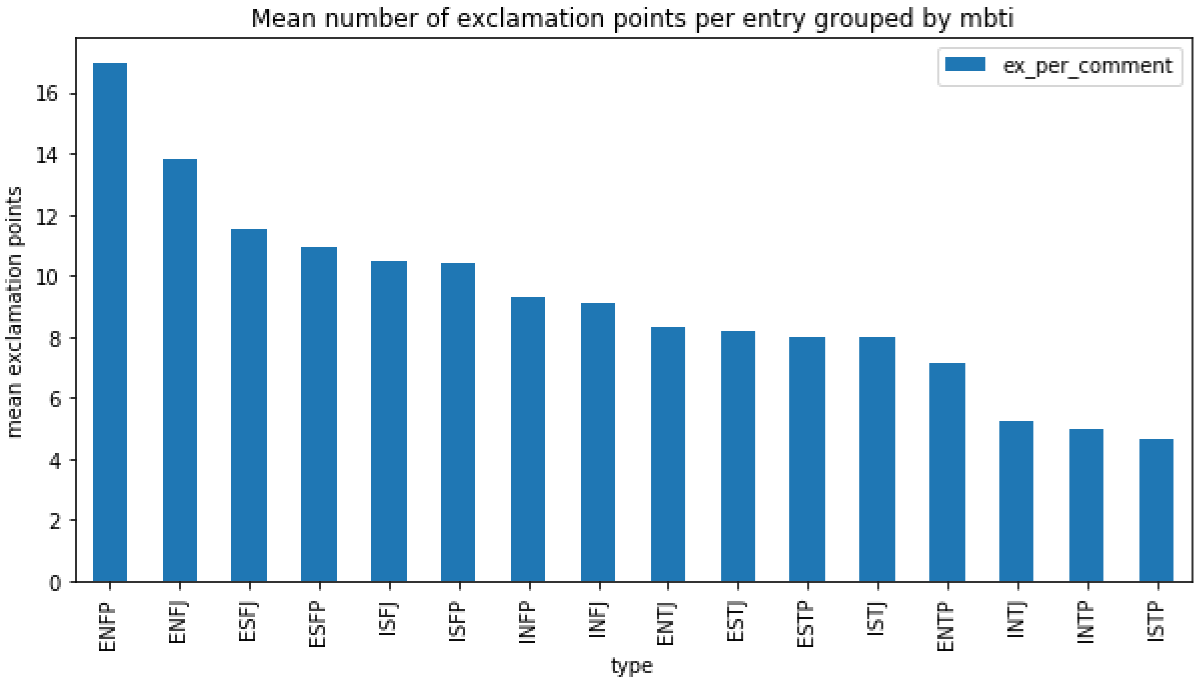


Figure 4. Mean exclamation points by each of the 16 personality classes.

6.2b Four Axes

Like in the EDA, it seemed prudent to explore these metrics from the perspective of the data sampled four times into the four axes. Remember that these axes contain overlapping features between axis, and cannot be compared beyond the confines of each axis.

Skipping over mean word count, which was covered in the results of the EDA, the next metric is variance, which no longer varied distinctly between axis ends (*figure 25*)

Mean number of HTTP links was again, hardly distinct, although interestingly the mean number of links for Extroverts over Introverts was 3.5 to 2.5 (all other axes fell between 3.0 and 3.5). It is always interesting to find features that appear to be independent from class to class (*figure 27*).

Mean number question marks was as unremarkable as the 16-classes analysis, with each axis falling between 11.5 and 12.5 links per document (*figure 28*).

Again, of all the metrics, Mean number of exclamation points per document was the most interesting metric. Of the four axes, E-I and F-T were the two with notable ranges, with 7.75 to 11.75 and 11 to 6 respectively (*figure 29*).

6.3 Machine Learning

Multinomial Naive Bayes, an often solid choice for text classification, struggled to describe this particular data set, as it was only able to score an accuracy of 25.0% fit to Sample 1 (the split data) and 40.4% fit to Sample 2 (the unsplit data).

Logistic Regression, another useful go to, fared much better in its own turn, as it was able to describe sample 1 and 2 with a score of 40.0% and 63.4% respectively.

The real winner at describing the data was a Support Vector Machine classifier trained using Stochastic Gradient Descent (SGDClassifier) which was able to classify all 16 classes with an average accuracy of 25% on Sample 1 and 67% on Sample 2 out of the box, which is already about 11x better than simple chance. This was selected as the model of choice for parameter tuning, which after selecting l1_ratio = 0.3, ‘hinge’ loss function, max iterations of 10, and penalty ‘l1’ by means of Parameter Cross-Validated Grid Search, was able to yield a 0.6% improvement in accuracy.

As mentioned above, a closer investigation of confusion matrix of the SGDClassifier trained on Sample 2 made it apparent that many of the classes were overlapping in features. Therefore, taking into account the previous classifier trainings, SGDClassifier was applied in four separate binary classification problems on Sample 2 divided by axis label. The best description of the way this was divided was that I made four versions of the data set for each of the four axes, in which each was only divided by two labels from corresponding opposite ends of each axis. *Figure 5* summaries this approaches performance on each sample.

Performance of SGDClassifier per Sample

Axis	Sample 1	Sample 2	Sample 3
EI	0.768	0.842	0.577
NS	0.864	0.875	0.579
FT	0.599	0.847	0.763
JP	0.612	0.799	0.568

Figure 5. Performance of SGDClassifier each sample by axis.

7. Analysis

7.1 EDA

From *figure 3* it was clear that only three of the four axes were statistically significant in difference in mean word count (the odd one out being the I-E axis). Given what information is readily available about each axis from online resources, it may be possible to reason that the distribution of personality type representation in the data was possibly due to certain behavior habits that could be attributed to those particular personality types. All excerpts in this section came from the MBTI Wikipedia page: <https://en.wikipedia.org/wiki/Myers–Briggs_Type_Indicator> (jump to “Four dichotomies”).

7.1a Extroversion VS Introversion

I initially assumed there would be more extroverts posting comments online, as I figured that an extroverted personality would be more comfortable with sharing with thoughts with strangers. My assumption was wrong, as there are nearly 3.5x the entries for introverts than for extroverts. Consider this passage from Wikipedia:

> People who prefer extraversion draw energy from action: they tend to act, then reflect, then act further. If they are inactive, their motivation tends to decline. To rebuild their energy, extroverts need breaks from time spent in reflection. Conversely, those who prefer introverts "expend" energy through action: they prefer to reflect, then act, then reflect again. To rebuild their energy, introverts need quiet time alone, away from activity.

If the internet is a place for *reflection* and not *action*, then that would explain why introverts are more active online according to this dataset. Looking at the word cloud (*figure 10*), there is some obvious overlap in word choice, but just as many differences. Something worth noting is that introverts appear to reference introverts when commenting, as opposed to extroverts who not only reference themselves but also introverts.

7.1b Intuition Vs. Sensing

This axis had the starkest difference in representation in the dataset (7477 intuition versus 1197 sensing, 6.25x more intuition entries than sensing). Their difference in mean word count could definitely have been affected by this representation skew. To understand the implications of this skew, consider this passage from Wikipedia concerning the I vs S axis:

> People who prefer sensing are more likely to trust information that is in the present, tangible, and concrete: that is, information that can be understood by the five senses... They prefer to look for details and facts. For them, the meaning is in the data. On the other hand, those who prefer intuition tend to trust information that is less dependent upon the senses, that can be associated with other information (either remembered or discovered by seeking a wider context or pattern).

Without philosophizing to much about the implications of the internet's effect on people's means of perceiving the world around them, I would consider this passage to be a decent explanation of the representation skew within this data set. The way that one interacts with and through the internet is fairly limited in sensory terms; we can read information, watch videos, and listen to audio. There is little to no physical aspect of the internet with which to interact. It makes sense to think then, that the internet would attract people who trust information as it exists in context with other information (from a technology primarily concerned with the disbursement of information). In other words, the internet is a provincial paradise for people on the intuition side of this axis, and a sensory void for those on the sensing end of the spectrum.

7.1c Thinking Vs Feeling

Here we see a more or less equal representation on both ends of the this particular axis. Perhaps there is no difference in the two that would govern different behavior online? Again, let's refer to what we already know from Wikipedia:

> Those who prefer thinking tend to decide things from a more detached standpoint, measuring the decision by what seems reasonable, logical, causal, consistent, and matching a given set of rules. Those who prefer feeling tend to come to decisions by associating or empathizing with the situation, looking at it 'from the inside' and weighing the situation to achieve, on balance, the greatest harmony, consensus and fit, considering the needs of the people involved.

So then it seems there isn't a difference between the two ends that would merit a contrast in their behavior online. However, this axis exhibits the biggest and most statistically significant difference in mean entry length. Based on this, one might assume that those on the thinking end of the axis, being more governed by logic and order, would post shorter comments. However, within the boundaries of what we can observe, that is already pushing the limits of can be assumed.

7.1d Judging Vs Perceiving

Not a whole lot going on here, besides a disproportionate representation between judging and perceiving. Is there an explanation for this (seemingly minor) distinction?

> According to Myers, judging types like to "have matters settled"... perceptive types prefer to "keep decisions open".

Perhaps then we might say that perceptive types are more likely to be open to and respond to discussions on the internet, as opposed to Judging types who might have already have their minds made up and thus find interaction on the web to be less appealing.

7.2 Inferential Statistics

Overall this step failed to produce any useful metrics for feature engineering, and didn’t produce many, if any, interesting results.

When investing the mean word count by all 16 classes, the Pearson correlation coefficient seemed to indicate that entry count weakly influences the average word count of that class (*figure 15*). This did not translate to the mean word count of each class within in axis, which was much more distinct between classes in three of the four axis (the E -I axis was the offender here) (*figure 25*). Upon looking at the standard deviation of each classes means within in axis, it was apparent that there would be a huge amount of overlap (each class had relatively similar standard deviations with around 141 words) which pretty much disqualifies mean word count as a metric since the greatest difference between any of the means within an axis was only 26.

The same was true for mean number of HTTP links from both perspectives (*figure 18*). The Pearson correlation coefficient between mean number of HTTP links and entry count indicated that high entry count also contributes slightly to a higher number of links (*figure 19*). From classes within axis (*figure 27*), the mean number of HTTP links had a high standard deviation across all classes. That, combined with differences that were indistinct caused me to abandon this metric all together.

Again, the same could be said of mean number of question marks (*figure 20*), which, while thoroughly unaffected by representation bias from a linear perspective (*r* = 0.06) (*figure 21*), yielded similarly indistinct differences and no notable grouping trends from a practical standpoint. This held true for comparison of classes within each axis as well (*figure 28*).

Mean number of exclamation marks was a little more interesting, and was most distinct among all 16 classes, and only weakly negatively correlated with entry count (*figure 23*). From an axes perspective, only the E-I and F-T axes showed distinct differences.

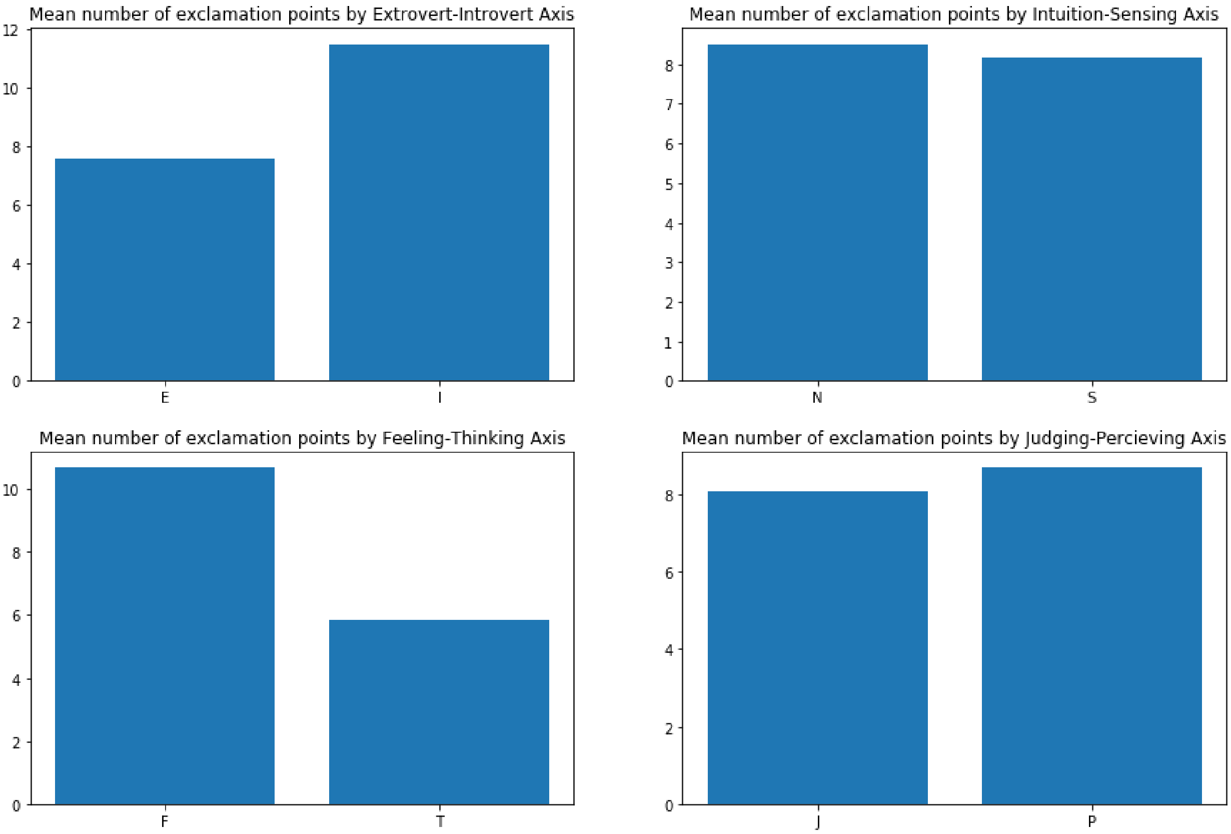


Figure 6. Mean number of exclamation points compared within each axis.

The E-I difference could be due to skewed entry counts, but the F-T axis was the most balanced of all the axes in entry count. The higher number of average number of exclamation points for Feeling reinforces the idea that Feeling people are more empathetic than the detached Thinking class. The large standard deviations made this metric not entirely practical for most uses, although it could be valuable as a feature to a corpus of particularly large document size (for reference, the mean length of all the documents in Sample 2 is 595.0 words).

7.3 Machine Learning

Across the board, all classifiers performed better with Sample 2 (the non-split data) than they did with Sample 1 (the split data). Having a greater feature set per document was a clear help to generalizing the classifier to test data, but the tradeoff was a substantially smaller set of total training documents.

When inspecting the metrics of the model attempting to describe all 16 classes, it became clear that there was not enough data to train the classifier to a practical level of accuracy. Indeed, from this point of view there was not a discernible way to evaluate the models ability to support the MBTI construct validity other than these simple metrics, especially since the model suffered so much from lack of support. It was these observations that drove me to select the four binary classifiers as my model of choice.

Using a single binary classifier per axis made much more sense because it much more closely mirrored the original scoring method of the test itself by comparing only two aspects of personality at a time (in this case from a text-feature perspective).

Once the model was trained and tested, I decided to take the probabilities of one class (within the axis) and graph it to see the distribution of the probabilities between the two classes (this works because the probability of class A is proportional to the probability of class B in a binary classification problem; 75% probability for class A is proportional to 25% probability for class B). My reasoning was that upon observation, a strong bimodal distribution between the classes would indicate that the classifier was confident in its classifications, thus supporting the MBTI personality construct. Conversely, a normal distribution of scores would reflect the poor retestibility of the original MBTI test (recall that the reason the MBTI was abandoned was because testees scores were normally distributed along each axis, making retest unreliable and thus calling into question the validity of the overall axis concept).

To account for the skewed data, I balanced the weighted classes, trading precision for recall. Afterall, it was more important to improve the classifications overall rather than have accurate true positives. The initial results on Sample 2 test data, while not perfect, were promising.

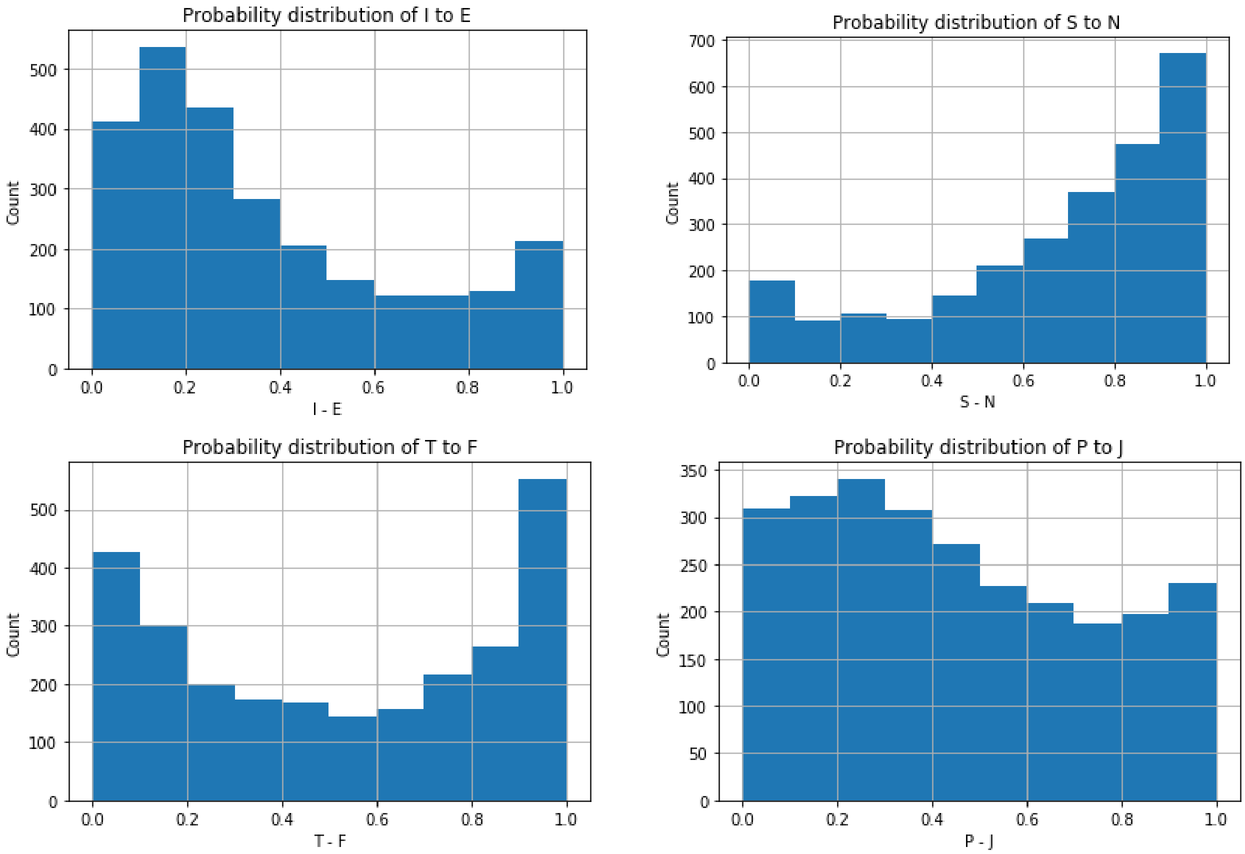


Figure 7. Test data probability distribution from classifier trained on Sample 2.

The lack of support in certain classes was affecting its ability to predict them, but with the balanced weighting there was a reasonable bimodal distribution along each axis. Pertaining to Sample 2, the results seemed to suggest support for MBTI as a construct of personality from a text based perspective. However, this success needs to be viewed in context to Sample 3 (recall that Sample 3 had all mentions of any of the 16 classes removed). These are the test data results after training the classifier with Sample 3.

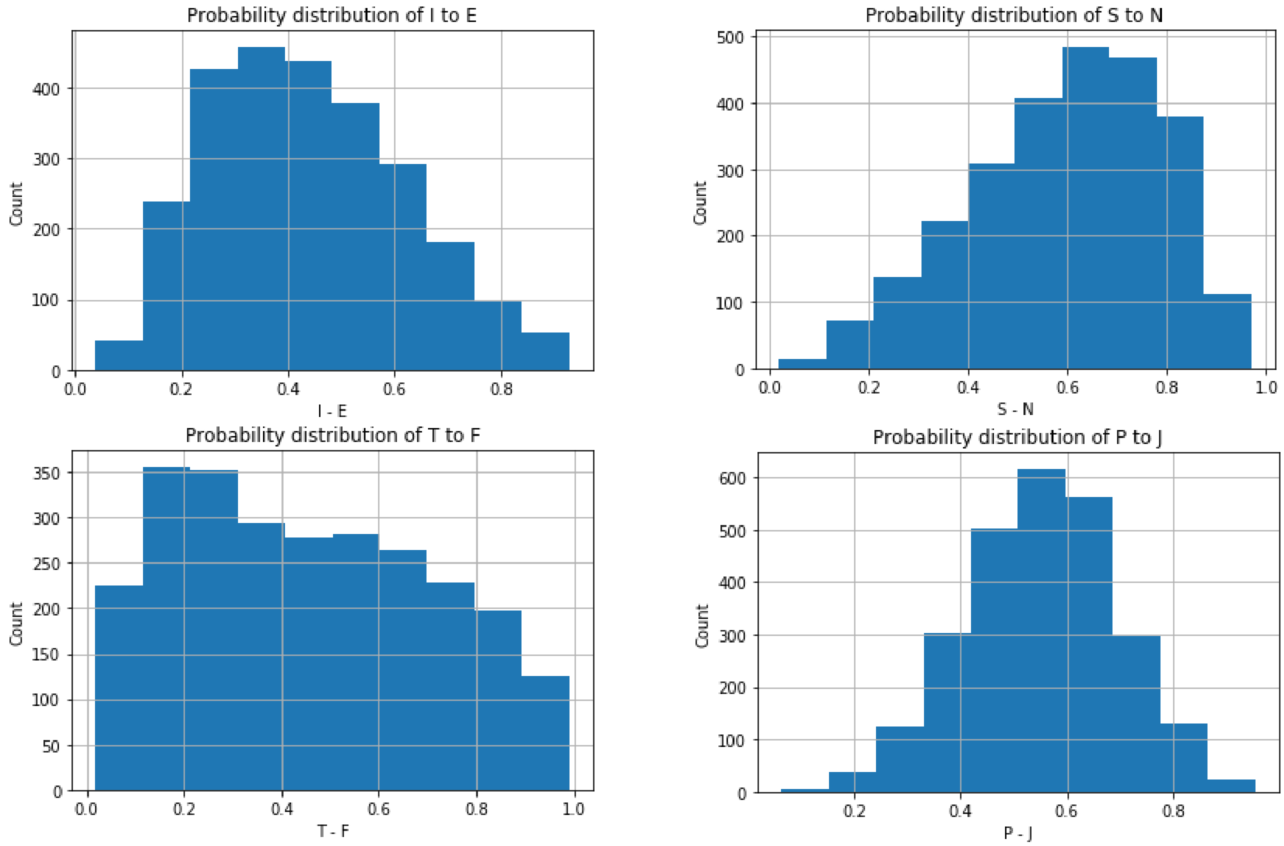


Figure 8. Test data probability distribution from classifier trained on Sample 3.

As we can see, the distributions are nearly perfect normal distributions. It would seem that the filtering out of all 16 class references removed one of the best distinctive features across all 8 classes in each of the 4 axis. To further test this observation, I fed the entirety of Sample 3 (not just the test samples) through the classifier trained on Sample 2 to observe the distribution.

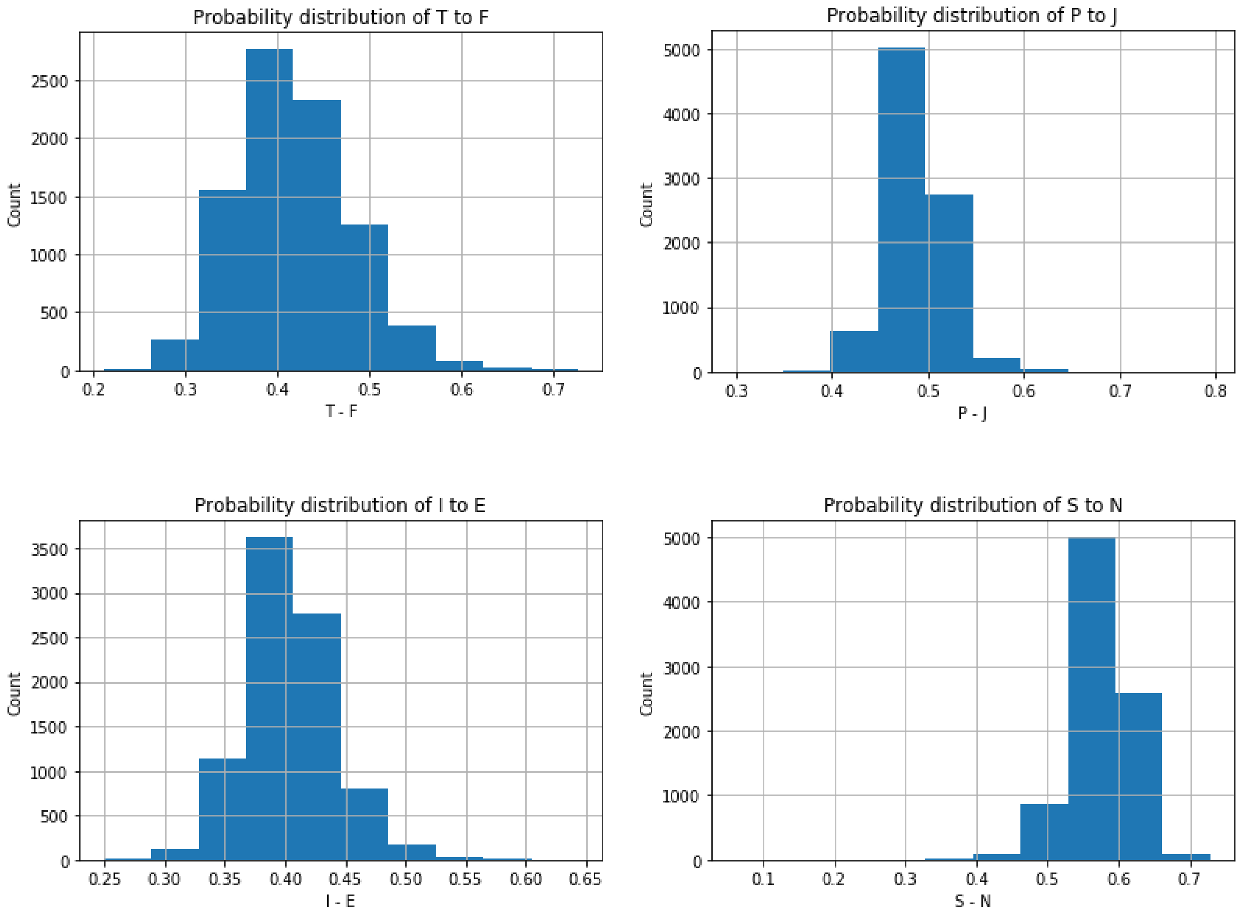


Figure 9. Probability distribution of Sample 2 classifier tested on Sample 3.

From this it is apparent that the classifier trained on Sample 2 did not generalize over in terms of predicting power on Sample 3, confirming my suspicion that the each of the 16 class references when present in the data are the most discriminative features of the personality classes.

8. Discussion

8.1 MBTI Personality Construct

So do these results mean that the MBTI personality construct is inherently flawed, as suspected? To a degree, yes. From the perspective of this data, there are a few conclusions that can be drawn. Primarily, that the context of where the data is being gathered is quite important, from a classification sense. This data was gathered from a forum where people discuss their personality types with one another openly. More importantly, they are aware of their personality type and are willing to share about it. Naturally this makes the data contextually reliant: in a place where people are aware of and constantly sharing about their personality types, they are also likely to adhere to certain behavioral patterns (in this case, language used in their generated text). In simpler words, whether or not someone identifies with their personality type (indicated again by the context of the data) and how it supposedly should influence their behavior seems to affect how well they can be classified according to the personality with which they identify.

While the inferential statistics did not prove to be very useful, the story told by the distribution of the data becomes more interesting now that we are aware that context can affect behavior (again, from a text-based perspective, not in a general sense). One must consider the fact that the reasons gathered within this report to explain the posting behavior of each personality class within each axis might be held as actual facts by the people who are actually doing the posting in the first place. It would be quite telling to observe this kind of distribution trend in data collected from another source (say, a large corpus of documents curated from another social media page dedicated to the MBTI). The bottom line is that, based on what was observed within this data set, one might say that belief in the contentions of the MBTI personality construct, again, might directly influence behavior.

In terms of Machine learning however, what these results might imply is that this model wouldn't generalize well to data collected from other areas (where the context is different). While we can only speculate, it would seem that that unless the data contains certain features likely to be observed within the context of people discussing the MBTI, it would be much less useful for classifying normal text collected from regular settings (like a twitter feed, or Facebook status updates). What this model can do, however, is identify individuals whose writings do contain those features. In other words, the value of this model lies in outlier detection! Recall that scores for each class along each axis were normally distributed in a setting with all 16 class references removed, but with the the references intact the model was very capable of separating out classes within axis definitively (think the bimodal distribution). Hypothetical user A (let's say from twitter) might post a large amount of MBTI related text, while hypothetical user B might post regular text not related to the MBTI. This model could easily pick out the user A out of a multitude of user Bs, showing us immediately who among a large group of users is likely to identify strongly with the MBTI.

8.2 Practical Application

Here we come to internal conflict: Usefulness vs Validity. Is the MBTI valid? Seems not, but it certainly holds practical value in describing the behavior of a subset of people. As an actual personality construct, the reliability of the MBTI is wobbly, but it clearly holds value to a large number of people to the point where knowledge of it might even dictate their behavior. In this context, the MBTI (and thus this model) holds usefulness.

An obvious application for this model is advertising. Advertising a product or service that targets some kind of subculture that might identify strongly with the MBTI could benefit greatly from being able to identify potential customers based on what kind of language they use in their social media posts. For example, a common target audience for book sales is introverts, who are considered to be quiet, private, and introspective (all a stereotype perhaps, but what is important is the extent to which they might identify with this stereotype, something this model is capable to detecting).This model could also be revealing about a particular target audience; do they respond to your product the way you expected them to based on the model’s predictions? What assumptions about the connection between their personality and purchasing habits proved to be true or false? The applications are extensive, especially since text data is so readily available from potential customers.

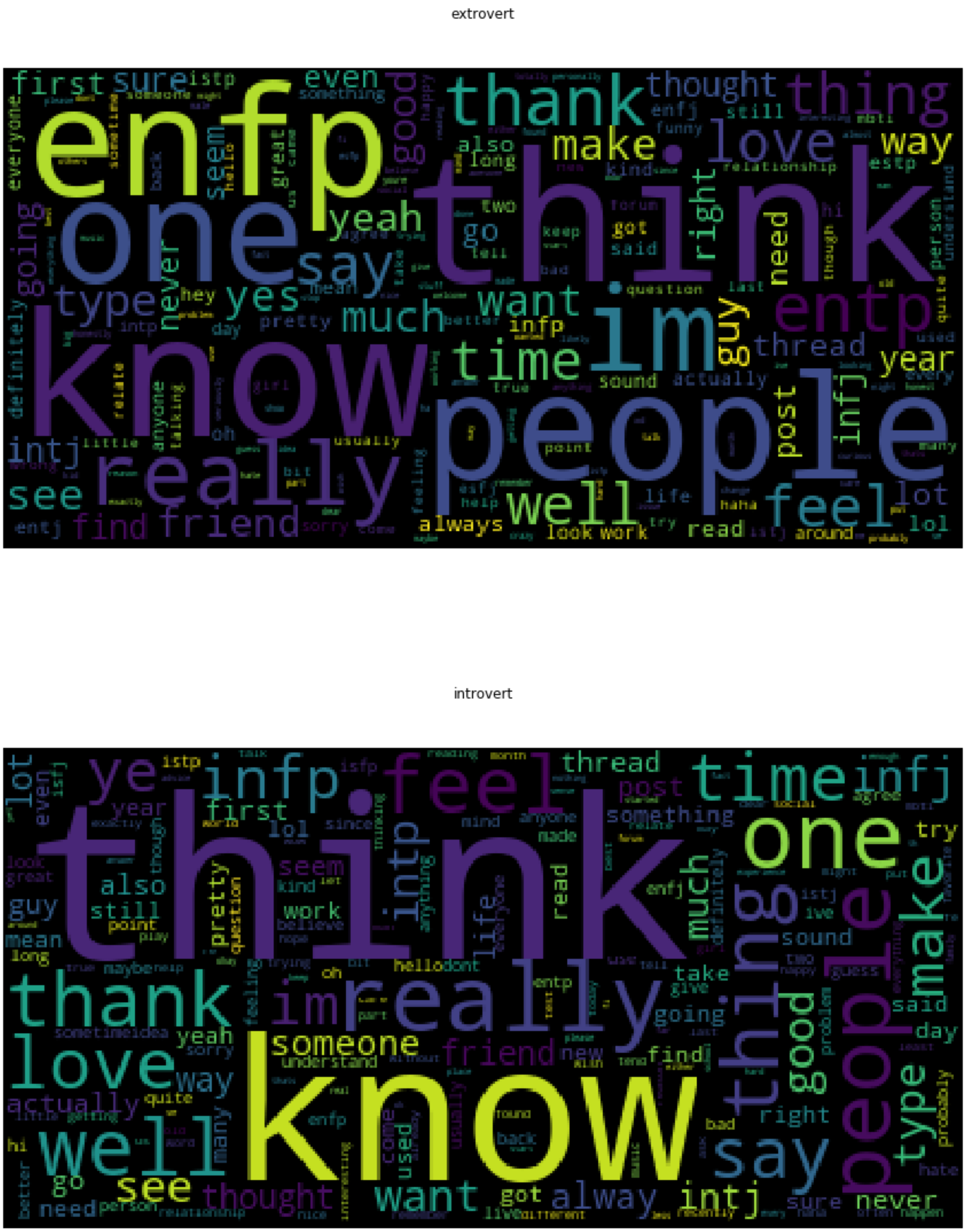
In the realm of quantifying impact of a piece of literature on a target audience, which in this case would be people who identify strongly with the MBTI, the only current metric to measure impact is views, shares, and comments. These are strong, ‘proof-in-the-pudding’ kinds of metrics, but are not available before the content is actually delivered. This makes content delivery a kind of trial and error process to find the right content to deliver to the target audience to get the best response. With this model in particular, a document could actually be scored on the probability that it represents particular personality class elements before even being presented. This model of course only would apply to text from the MBTI perspective, but it does hold implications for the idea of modeling a target audience’s response to a document and evaluating a document before hand from a purely text-based perspective.

An extension of this concept could also apply to the hiring process many companies are regularly forced to sludge though. One of the most time consuming processes of hiring, though arguably the most important, is applicant screening. This model was applied to text collected from social media, so it could in theory be applied to text collected from online that was generated by a person of interest, although it would only be useful in picking out individuals who strongly identify with the MBTI. In this sense alone the model isn’t entirely useful, especially since there is no way of knowing whether the MBTI personality classes can represent qualities searched for in a good hire, but again it does hold implications for modeling successful hires using text data. This text in question could be collected from sites that persons of interest post on (Linkedin, Twitter, etc.) but I think it holds more interesting potential in resume screening. By using the same methods outlined in this study, one could explore the idea of modeling company-specific successful hires using the text provided in their resumes (or evaluate the notion that there is even a concept of specific text features appearing in a resume that would improve chances of successful hiring). That however is an entirely different study, for which this one might serve as inspiration and, to a limited extent, a proof-of-concept.

9. Conclusion

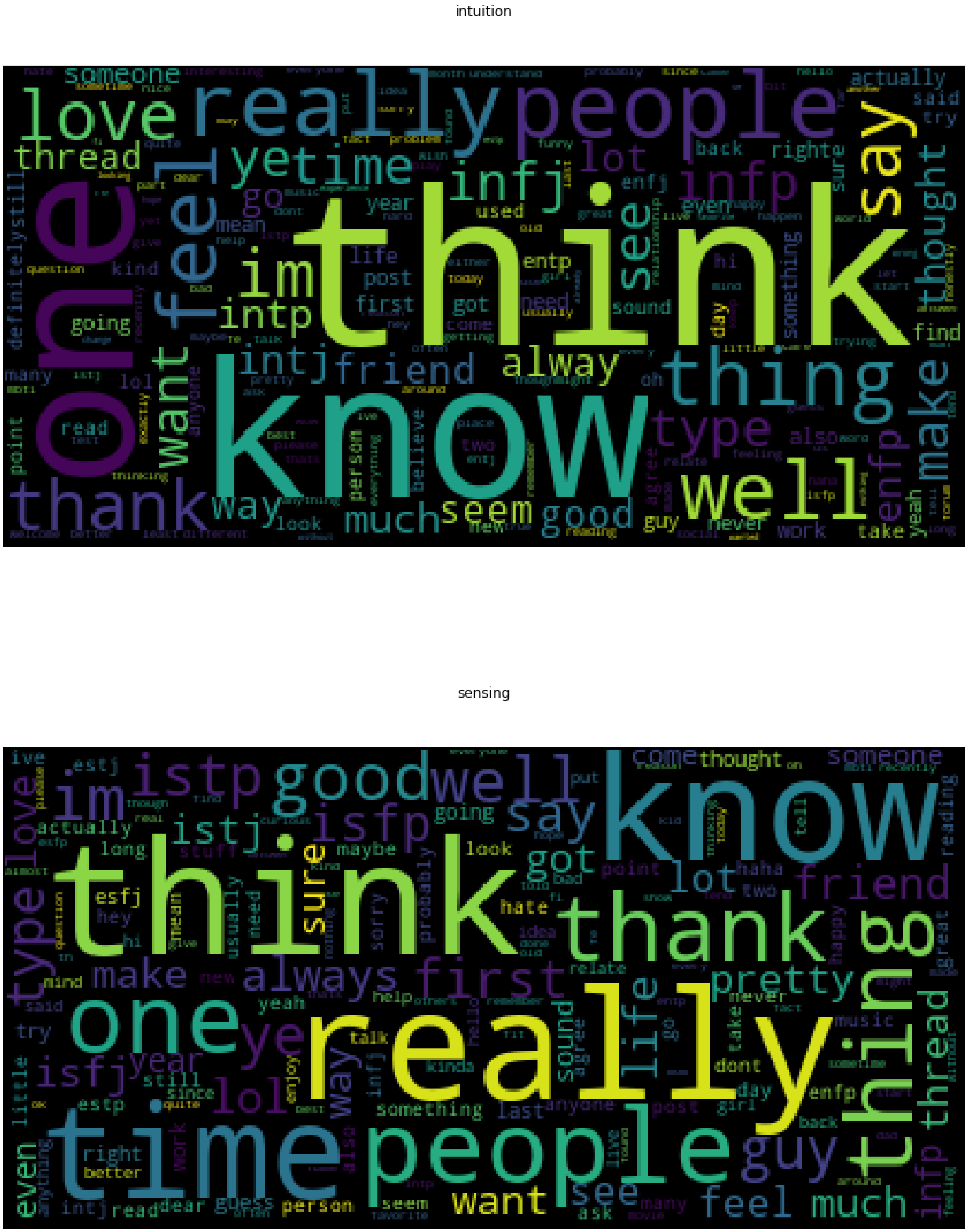
Hopefully then this analysis can serve as a strong reminder of the power of context, while also serving as an encouragement to consider the value of text features applied to classification problems. Text is something that is readily available in multitudes of observations, and is so often discarded as being too complex to be useful. From what we can see in this applied analysis, context can add a powerful element to what might otherwise be a useless bag of words, and when considered carefully enough can actually add significant value to what otherwise might be considered the noisiest kind of data around.

10. Appendix


extrovert

introvert

Figure 10. Word cloud of top words appearing in Introvert - Extrovert axis.



intuition

sensing

Figure 11. Word cloud of top words appearing in Intuition - Sensing axis.

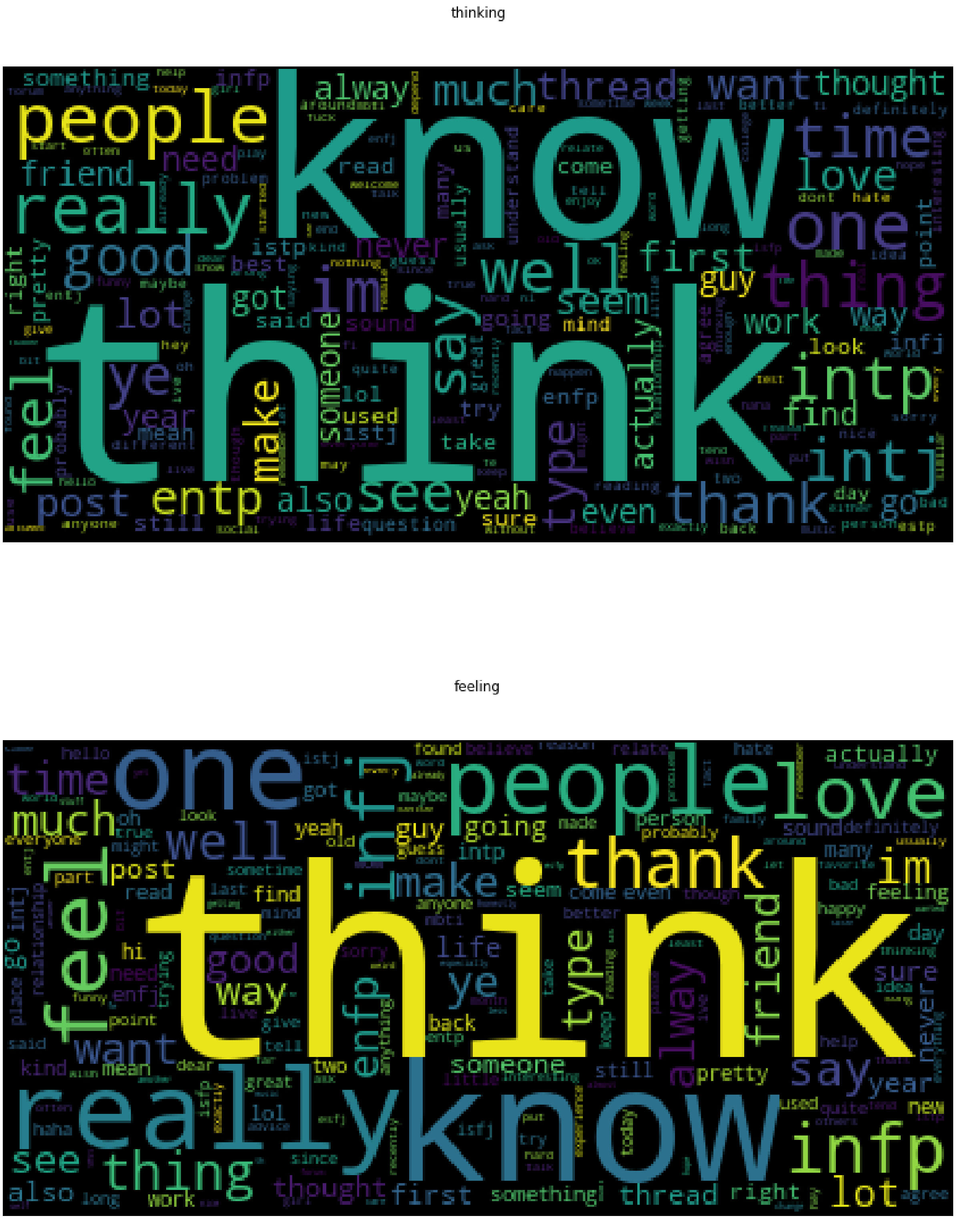


Figure 12. Word cloud of top words appearing in Thinking - Feeling axis.

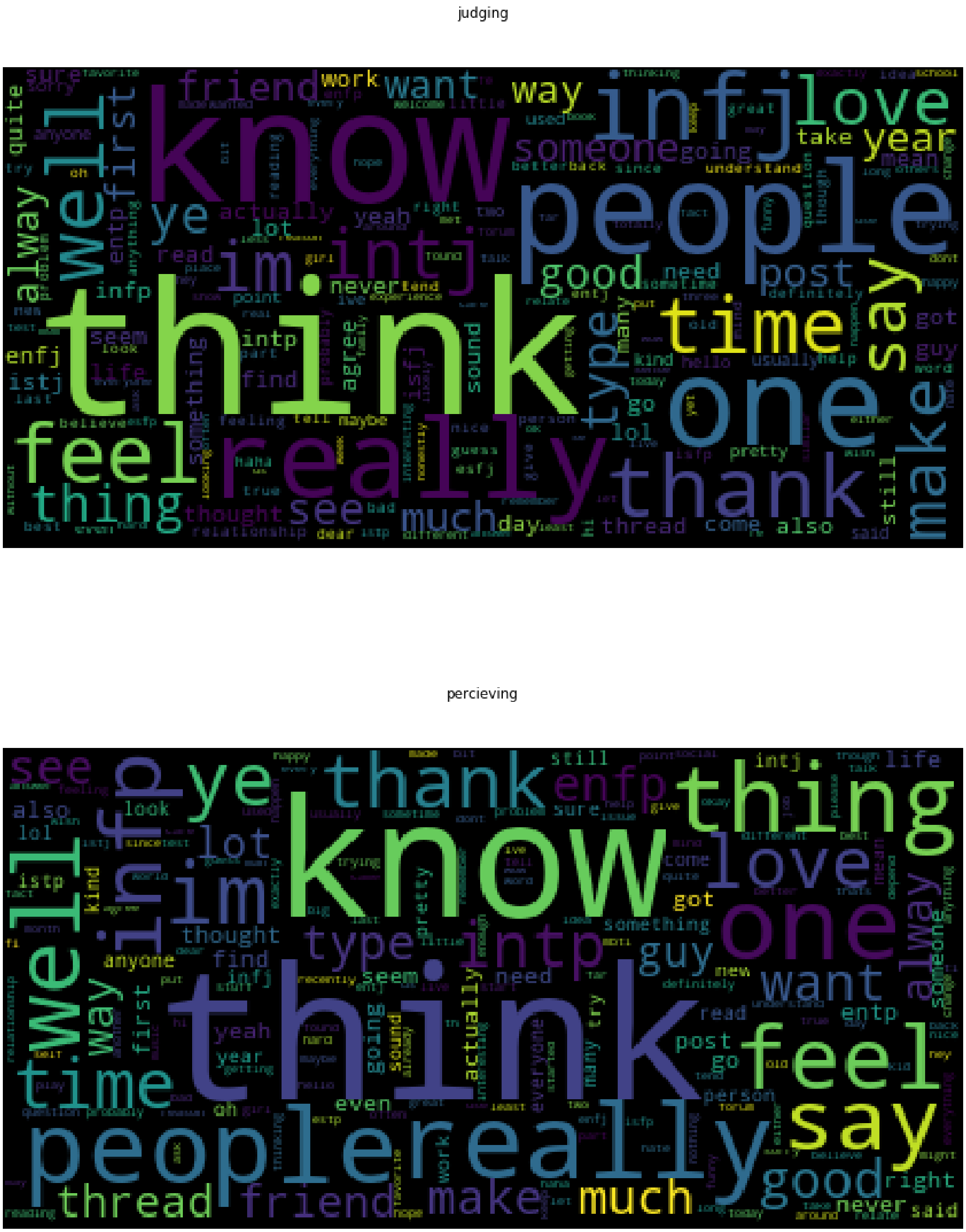


Figure 13. Word cloud of top words appearing in Judging - Perceiving axis.

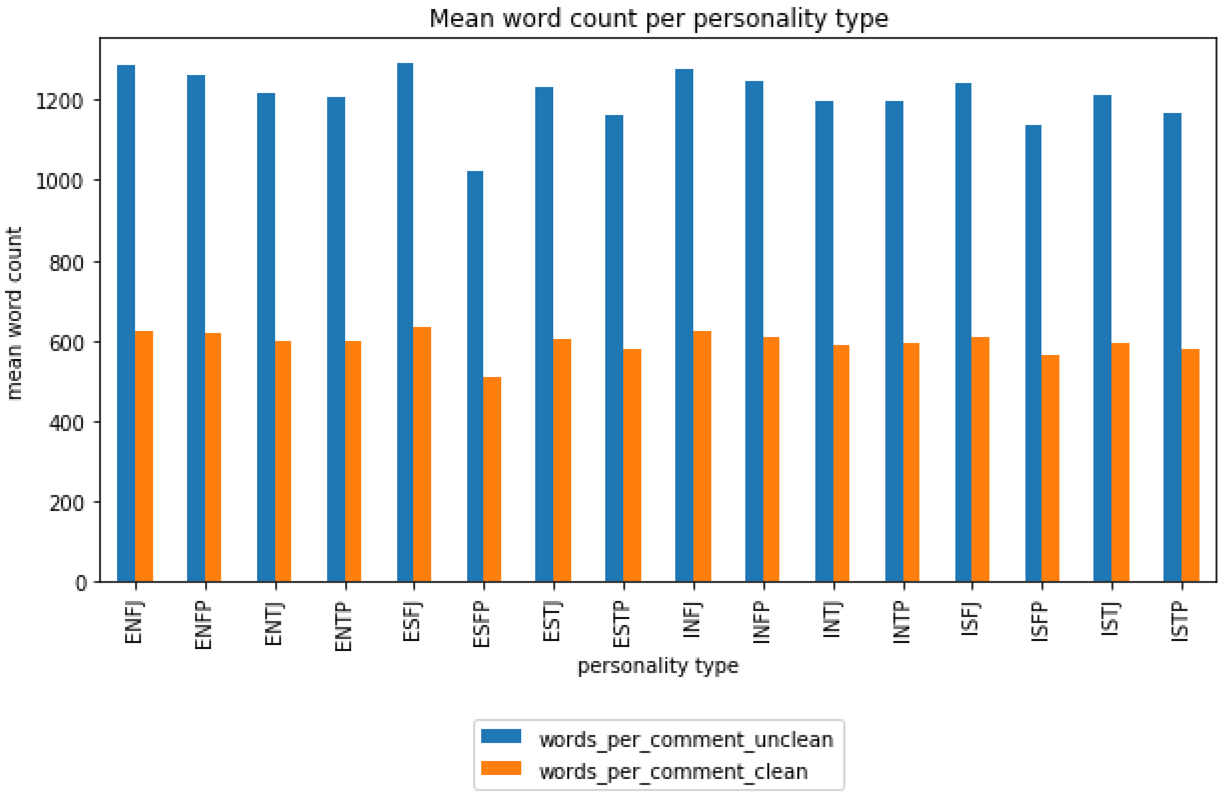


Figure 14. Mean word count per personality type appearing in data set.

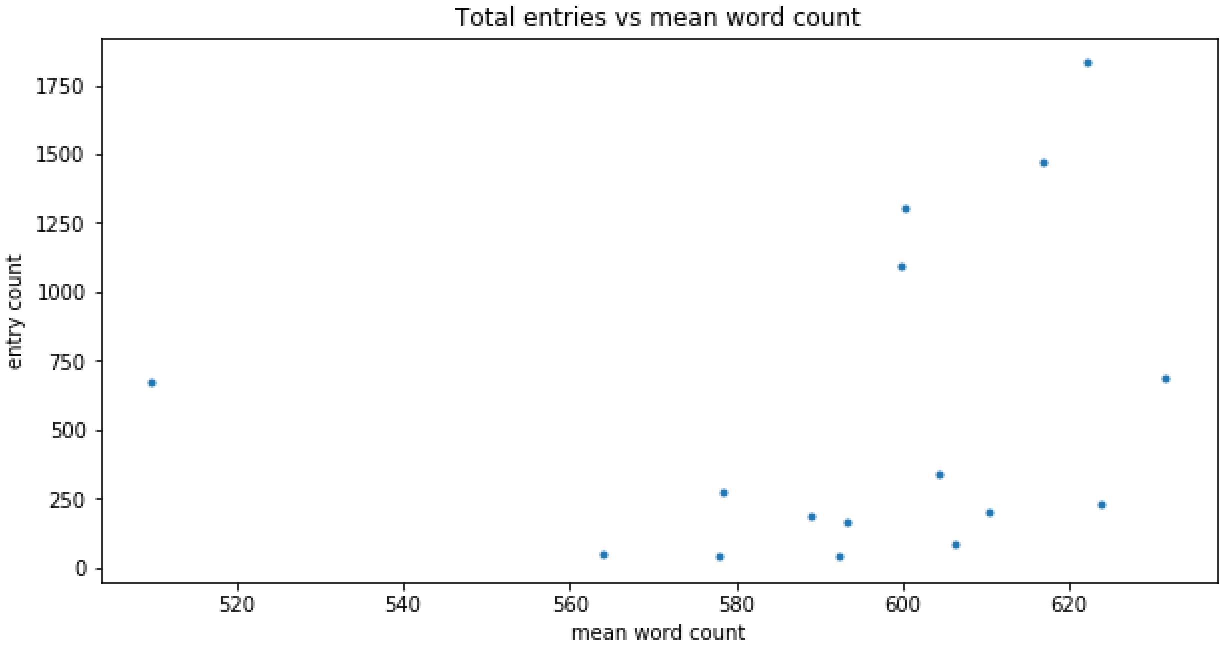


Figure 15. Correlation between mean word count and entry count per personality type.

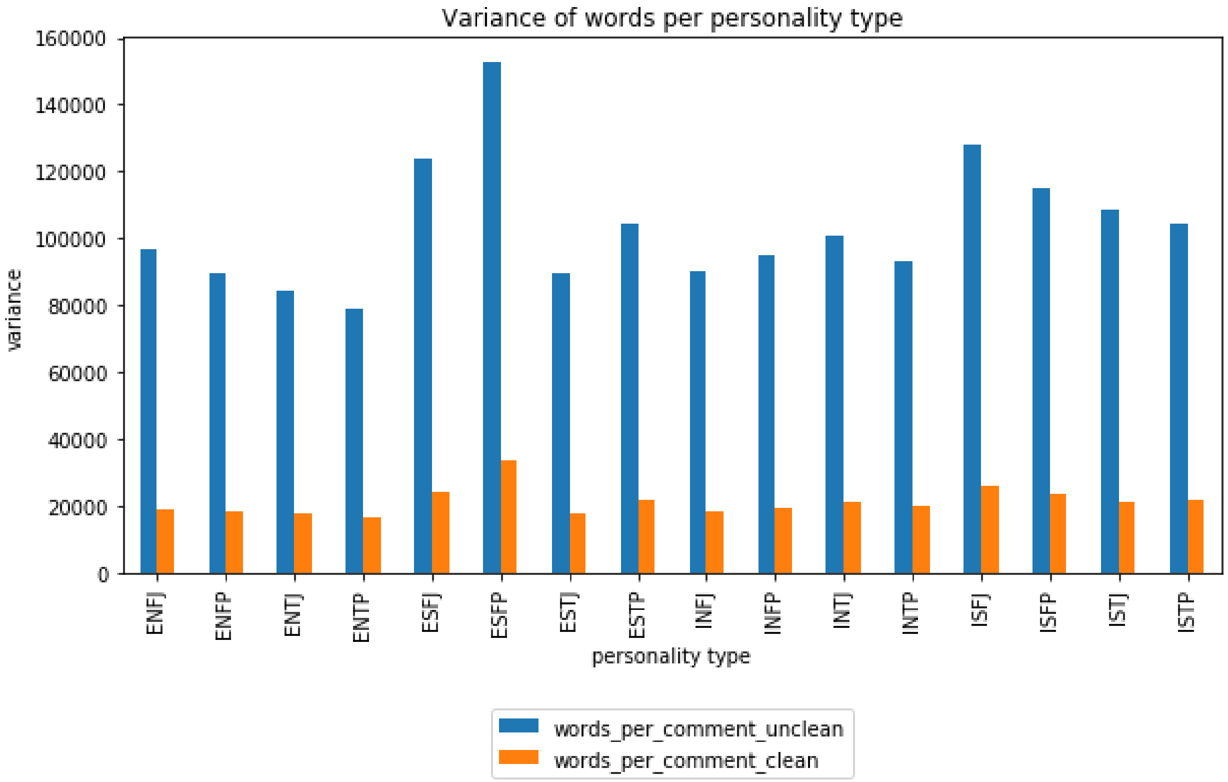


Figure 16. Variance of word count per personality type appearing in data set.

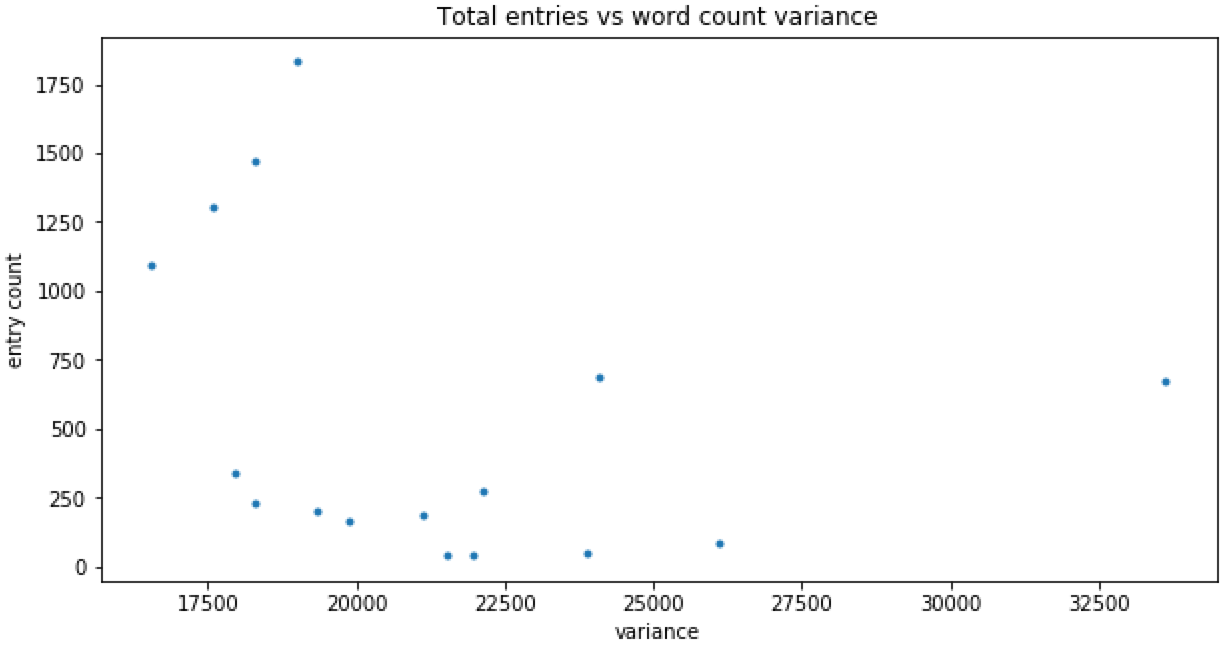


Figure 17. Correlation between word count variance and entry count per personality type.

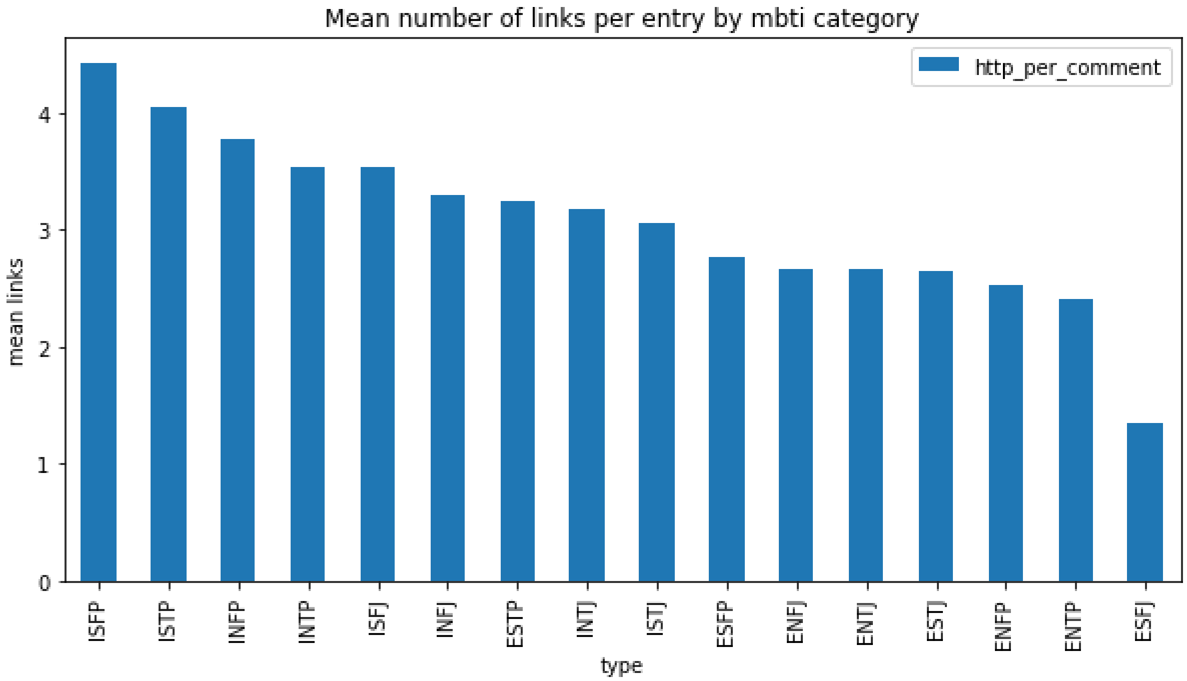


Figure 18. Mean number of links per personality type appearing in data set.

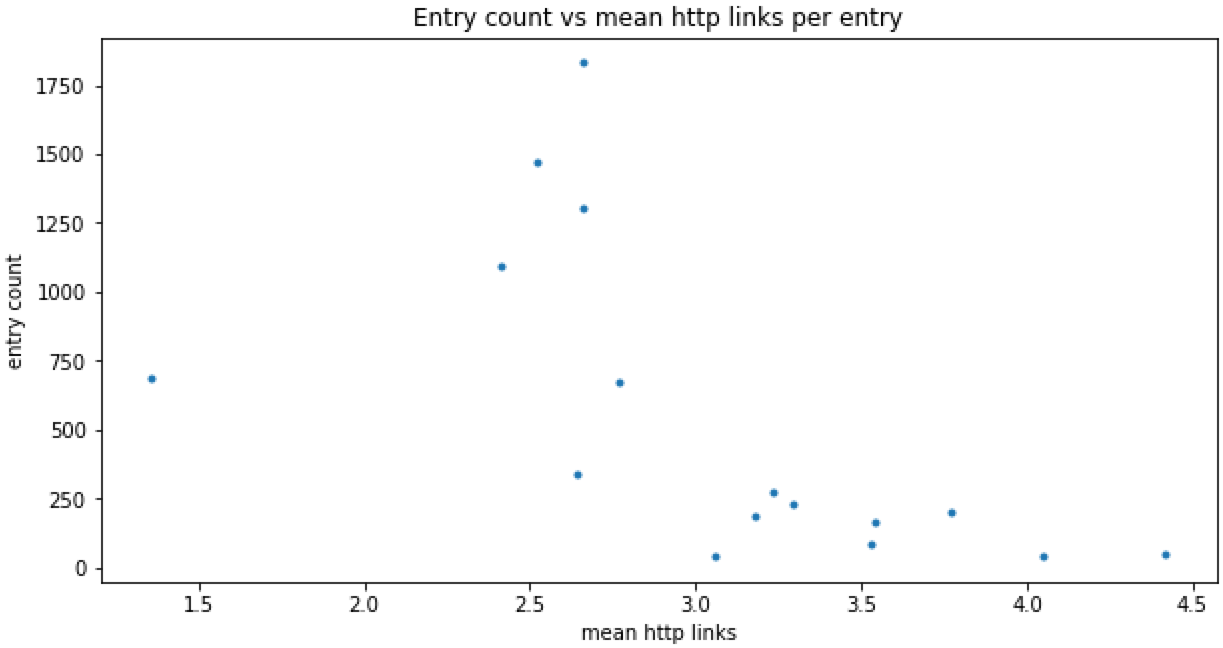


Figure 19. Correlation between mean number of links and entry count per personality type.

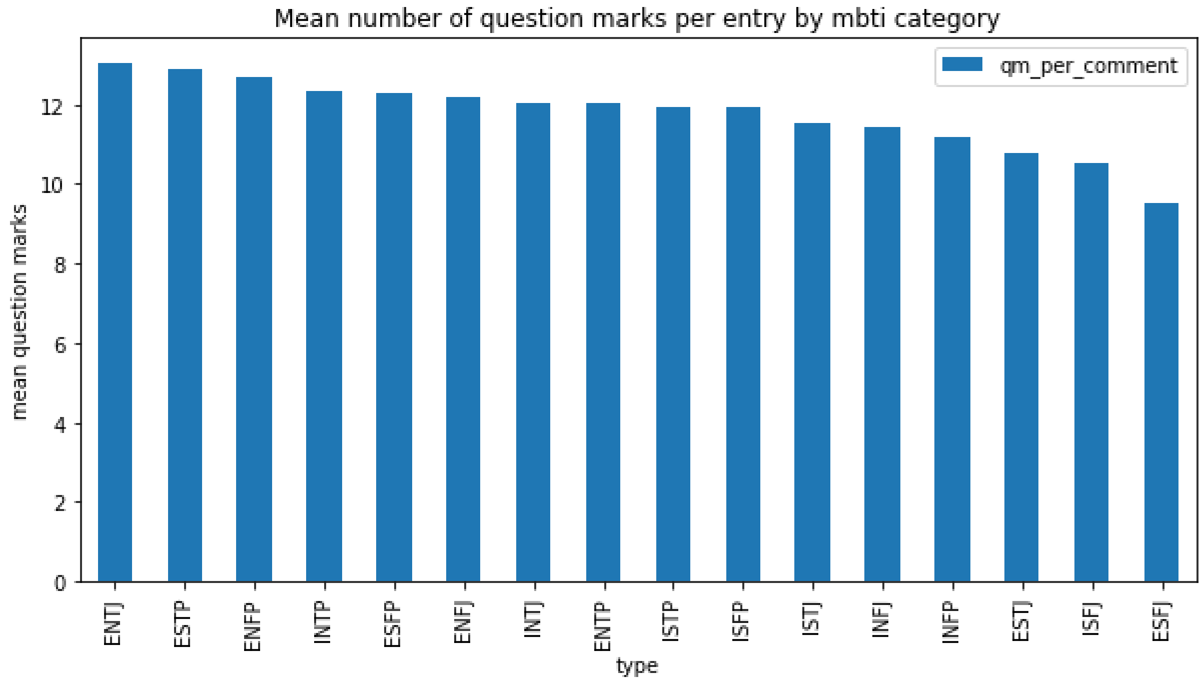


Figure 20. Mean number of question marks per personality type appearing in data set.

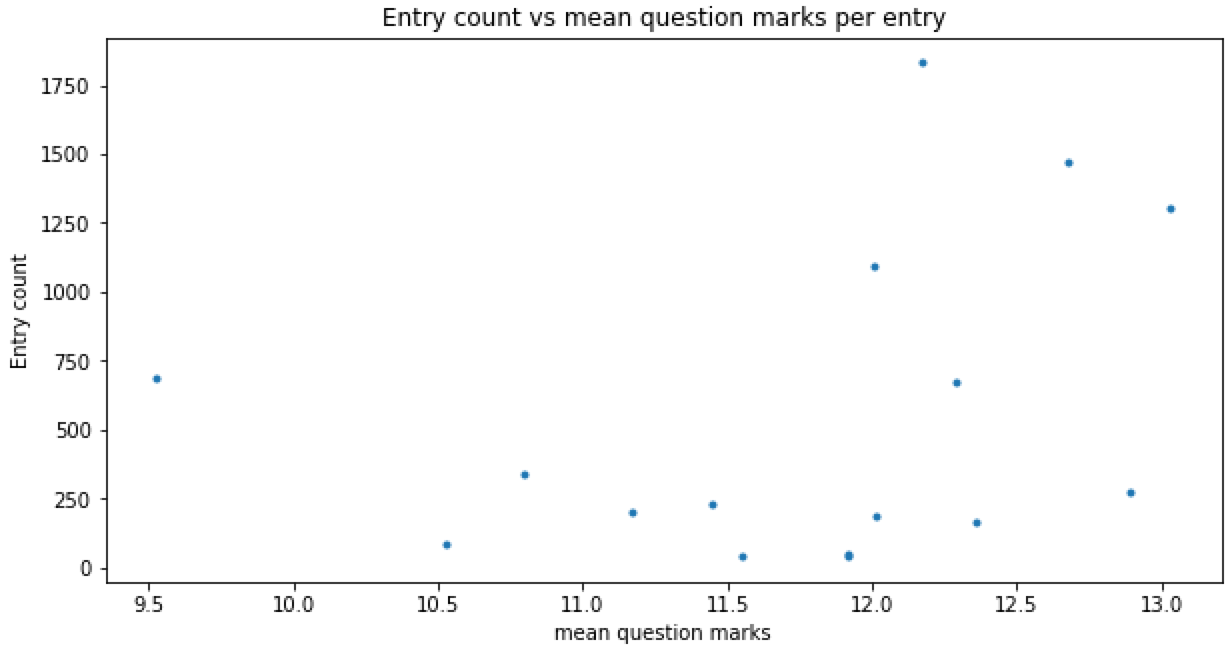


Figure 21. Correlation between mean number of question marks and entry count per personality type.

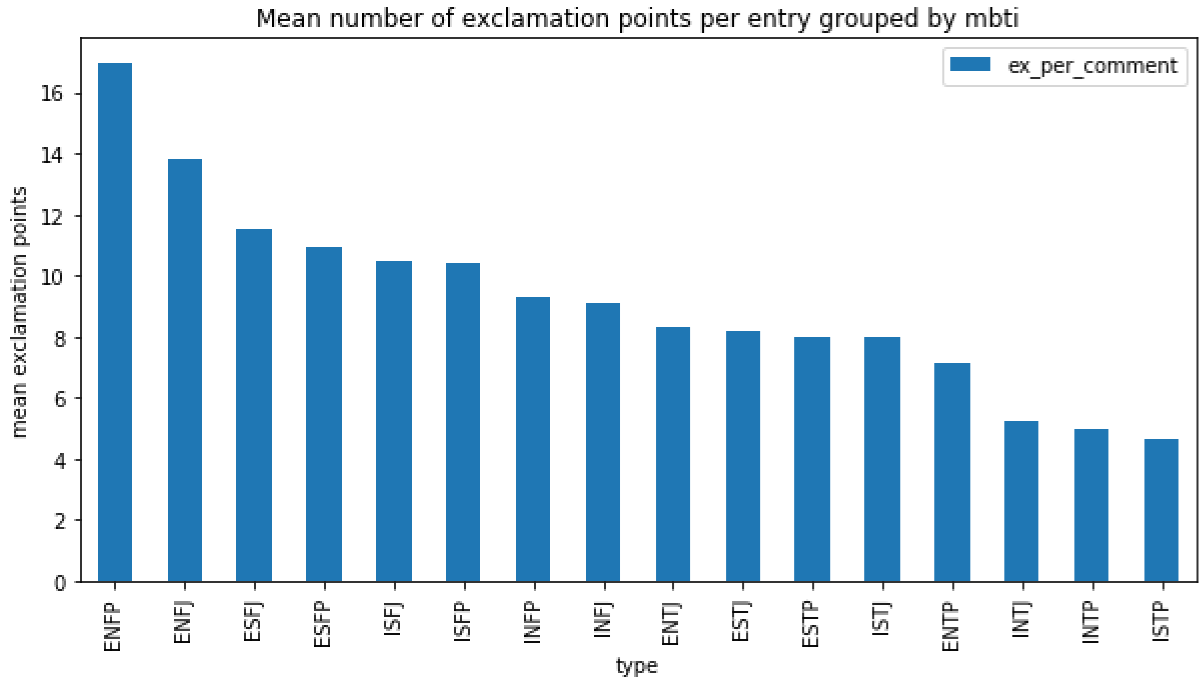


Figure 22. Mean number of exclamation points per personality type appearing in data set.

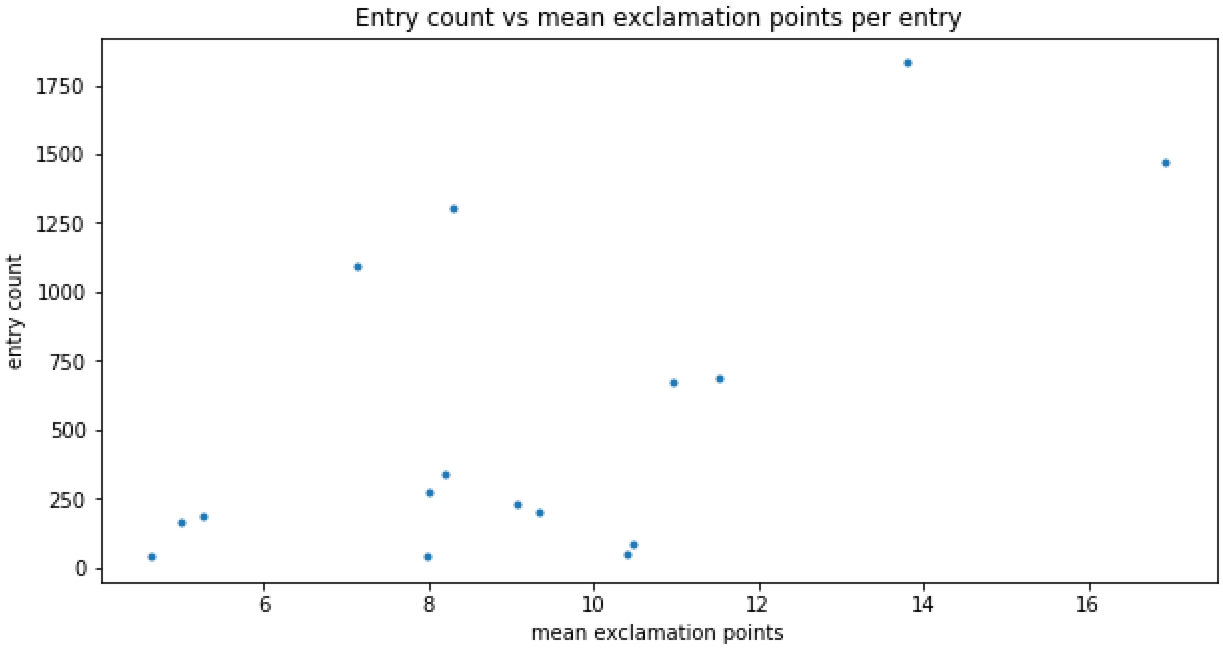


Figure 23. Correlation between mean number of exclamation points and entry count per personality type.

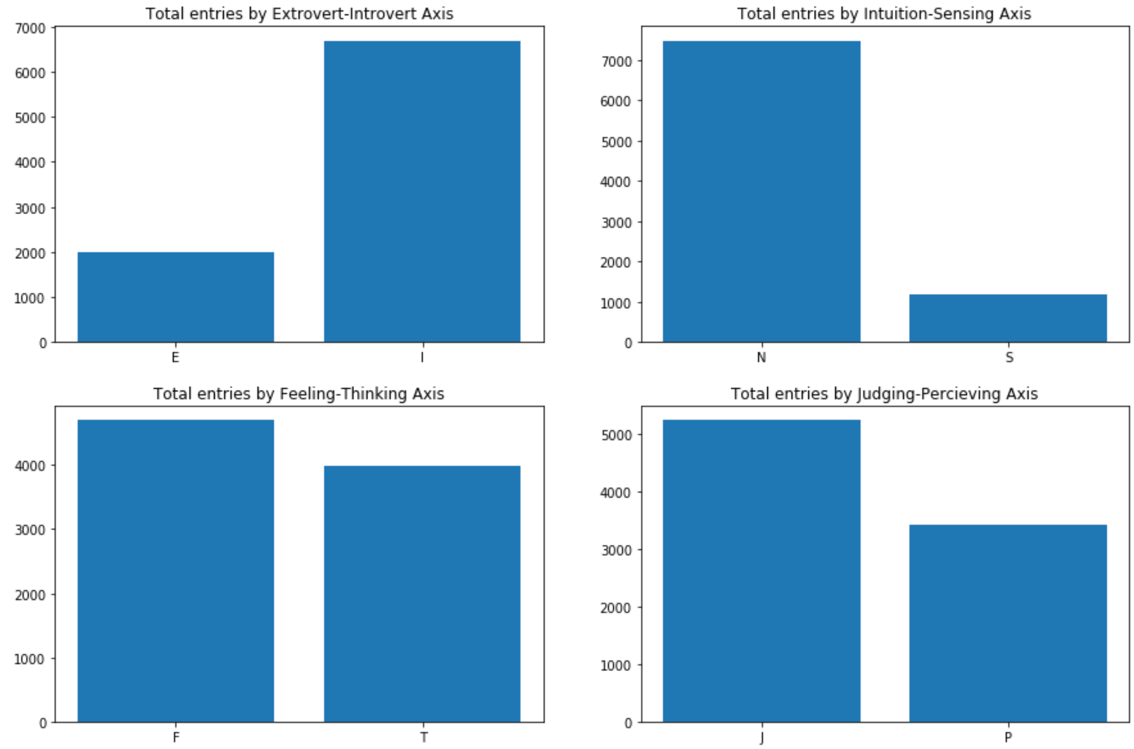


Figure 24. Total entry count of each personality class within each axis.

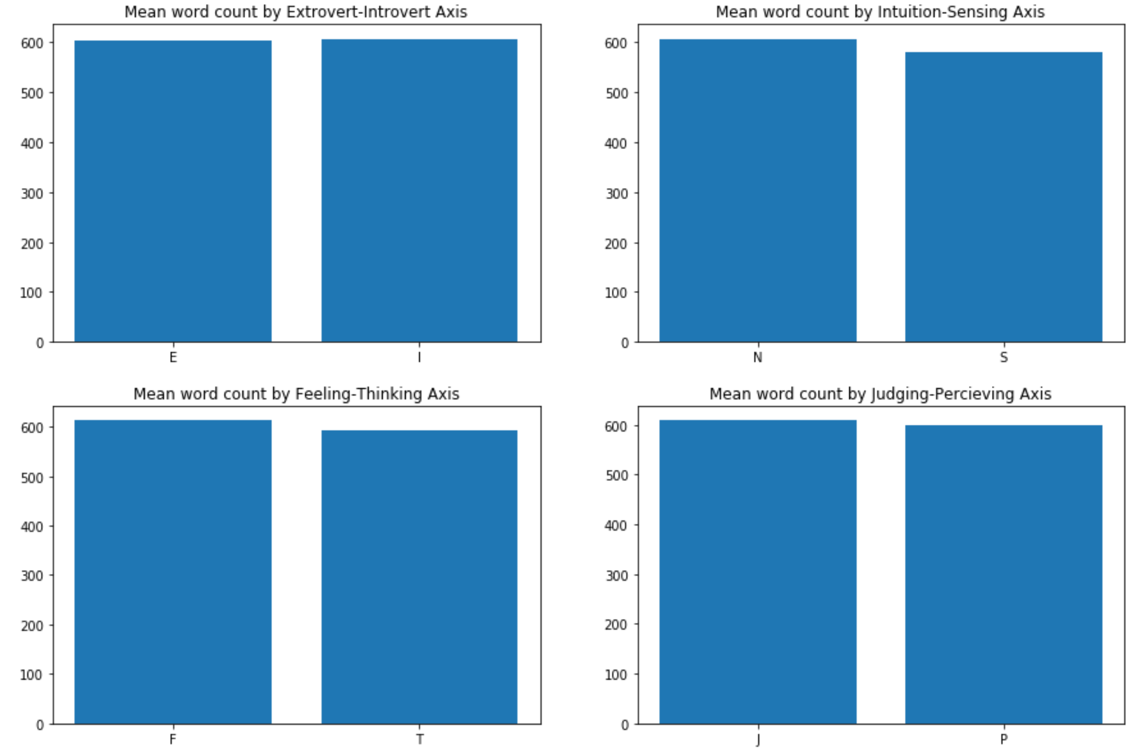


Figure 25. Mean word count of each personality class within each axis.

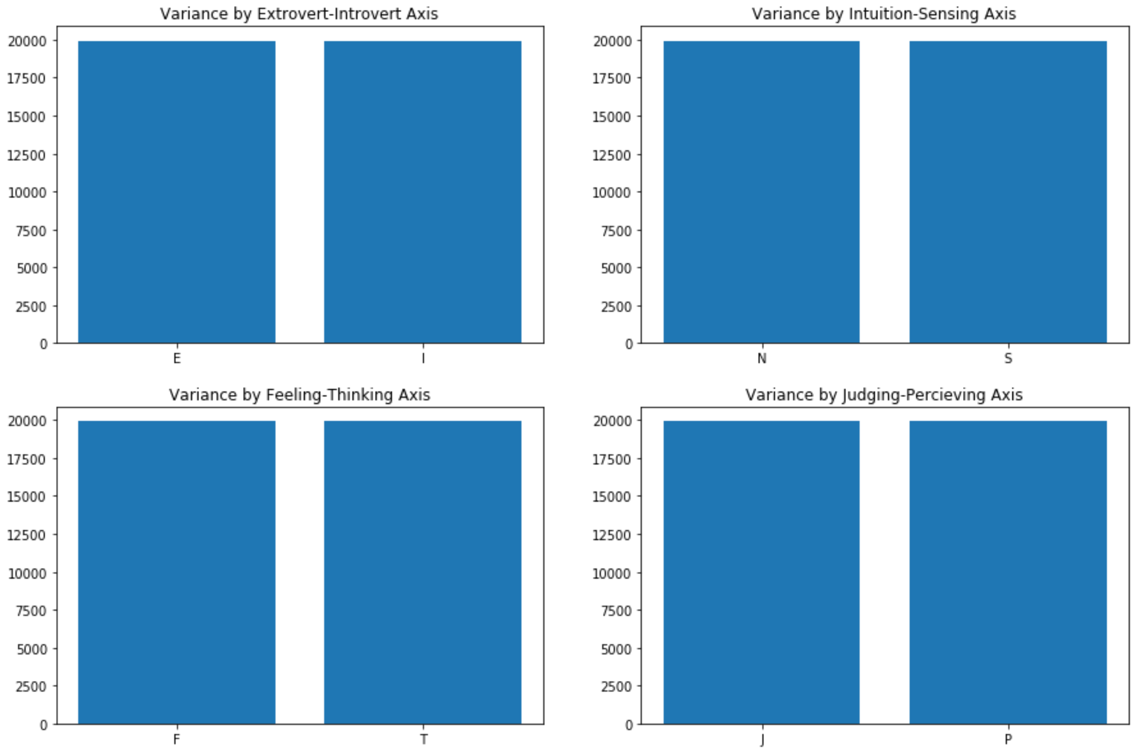


Figure 26. Variance of word count of each personality class within each axis.

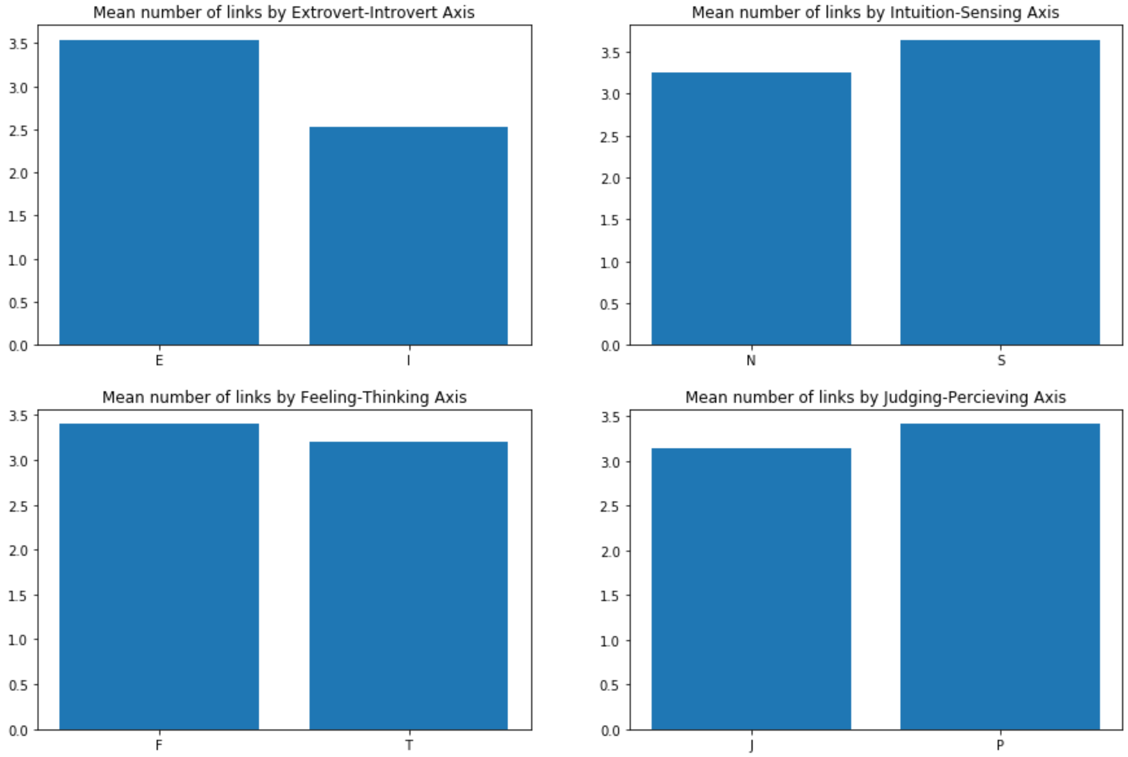


Figure 27. Mean number of links in each personality class per axis.

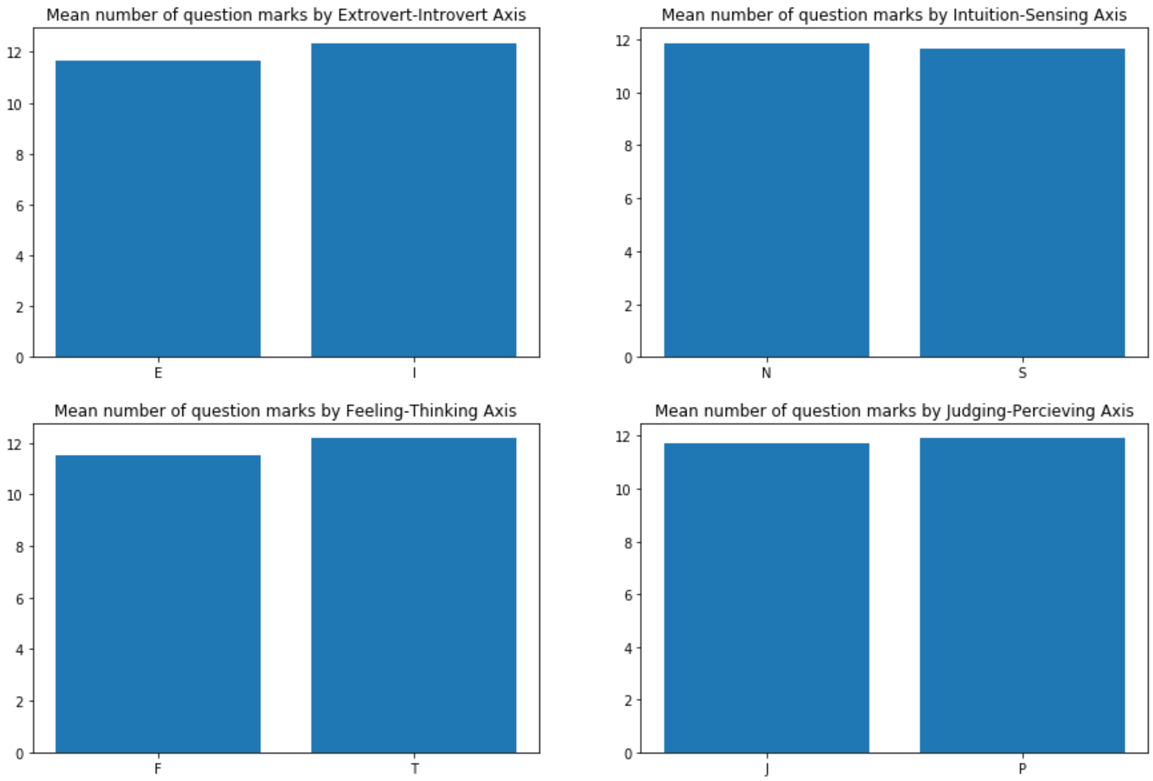


Figure 28. Mean number of question marks of each personality class within each axis.

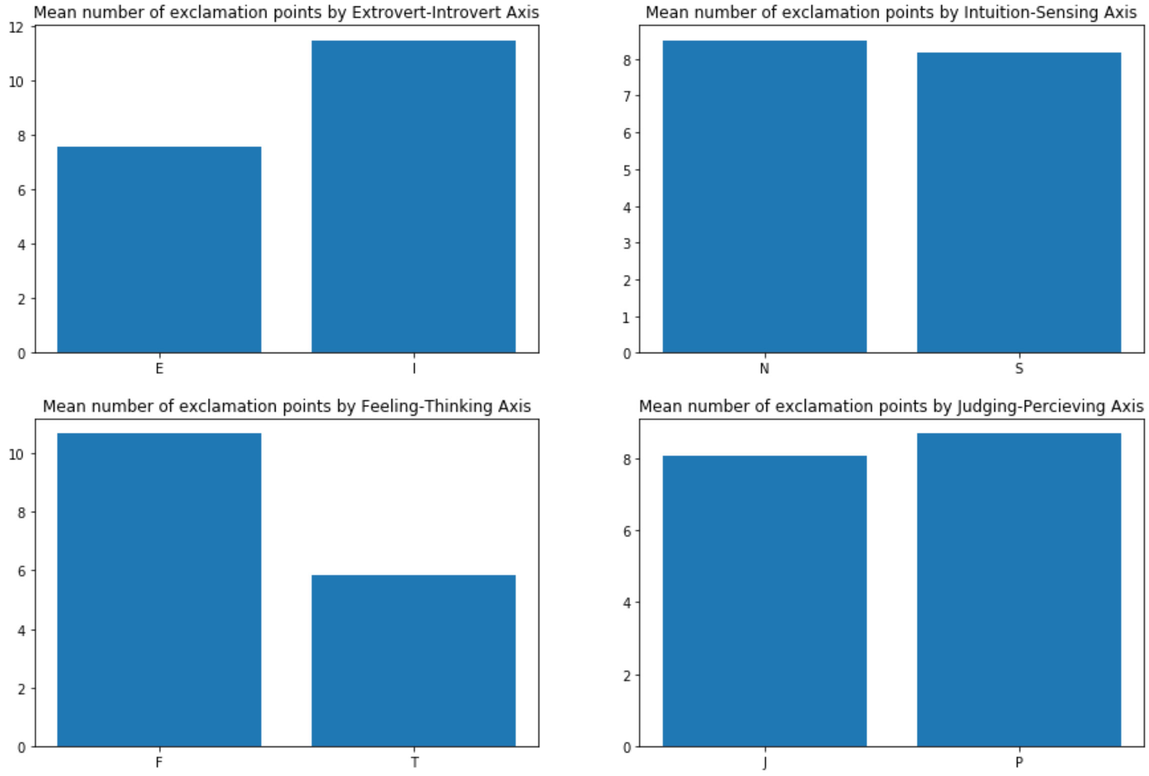


Figure 29. Mean number of exclamation points in each personality class per axis.