

Automated IQ Estimation from Writing Samples

Austin Hendrix, Roman Yampolskiy

University of Louisville

Louisville, KY 40208

austin.hendrix@louisville.edu, roman.yampolskiy@louisville.edu

Abstract

The primary focus of this research is to introduce a method of measuring an individual's IQ by analyzing the vocabulary in said individual's writing. In this paper, we show that the ratio of SAT words in a dataset of writing samples is roughly normally distributed, though with an obvious left skew. We go on to show a method that can be used to calculate an individual's IQ with this ratio and provide samples with measured accuracy. The conclusion suggests ways to increase accuracy in order to further develop the research along with applications of doing so.

1 Introduction

Stylometry is the statistical analysis of differences in literature between authors (Franking, 1988). As early as 1880, the study of stylometry has been used as a method of authorship identification on disputed texts. With the development of computers and automation techniques, stylometric analysis has become easier. An early example of software defined stylometry was used to identify the author of the disputed papers amongst the "Federalist Papers" (Tweedie, Singh, Holmes; 1996). This work demonstrated that stylometric analysis using automation is, at least in this application, able to draw similar conclusions about the authorship of these papers as previous work on the subject. In recent years, stylometry has taken on a broad range of applications. More specifically, stylometry has been used in the identification of chat bots (Ali, Hindi, Yampolskiy; 2011). Further research was done to show that when a chat bot changes behavior over time, the stylometry approach becomes more difficult (Ali, Schaeffer, Yampolskiy; 2012). In addition, it has been demonstrated that stylometric author identification processes can be used on a single author that is capable of writing in multiple languages. (Ali, Yampolskiy, 2014). This is significant in that it

demonstrates certain writing trends are independent of the authors' language and are therefore likely stronger candidates for comparing authors that write in different languages.

As there is no true scientific measurement that is currently used to quantify someone's intelligence, many different measurements have been used. Intelligence tests have often been a common way to determine an individual's intelligence relative to others. There have been many negative and controversial opinions on these tests, yet experts still agree on their overall usefulness (Snyderman, Rothman; 1989). Further studies have shown that a standard intelligence test provides the best single, reliable predictor of academic aptitude (Bullerdiek, 1985). One popular example of standard intelligence tests measures an individual's Intelligence Quotient (IQ). The assumption behind this system of measurement is that if a large sample of IQs are mapped together, the distribution will be normal. It has been shown that there are issues with the structure and quality of the standard IQ test (Lawler, 1977). Still, the IQ test can be a useful way for individuals to compare intelligence. For this paper, we will act under the assumption that an individual's IQ score relates directly to their true intelligence level.

This preliminary research project is focused on exploring whether an individual's IQ can be determined by using software defined stylometry. The novelty of this process is that it is not centered around author identification. Instead, stylometry will be used to determine the relative writing quality of a known author. The process will involve analyzing an attribute of a known author's writing to determine said author's IQ. There are multiple attributes of writing that are potential candidates for this application. For the beginning of this research, we will focus on the individual in question's vocabulary. Other research has been done to discuss other attributes with possible merit. These

attributes include, but are not limited to, word-length, syllables, sentence-length, and distribution of parts of speech (Holmes, 1994).

2 Collegiate Word Ratio

To determine an individual's IQ based on their vocabulary, a quantitative way to measure the quality of their vocabulary is necessary. For the purposes of this project, we will define a "Collegiate Word" as a word the SAT considers a part of strong vocabulary usage.¹ The College Word Ratio (CWR), which we will refer to through this paper, is therefore defined as:

$$\text{Collegiate Word Ratio} = \frac{\text{Collegiate Word Count}}{\text{Total Word Count}}$$

The CWR of each sample will be measured by software and then compared to the rest of the samples to determine its relative quality by use of a distribution. A pseudo-code for calculating the CWR of a sample is shown in Figure 1.

```
for SampleWord in Sample:
    for CollegiateWord in CollegiateWordList:
        if SampleWord == CollegiateWord:
            CollegiateWordCount++

CollegiateWordRatio = (CollegiateWordCount /
SampleWordCount)
```

Figure 2: Calculating CWR Pseudo-code

Now that we have a clearly defined a process for calculating the CWR of a sample, we need to execute this software on a large dataset. An ideal dataset would consist of writing samples by many randomly selected individuals. Along with this, each writing sample would represent each individual's average writing ability. As such a dataset was not available to the authors of this project, another source had to be found.

The Common Crawl is a corpus containing raw web page data, extracted metadata and text extractions.² The text extractions from this corpus contain the raw text taken directly from websites. We are acting under the assumptions that the text extractions are all written by humans and likely contain that individual's average writing. To help increase the accuracy of results under this assumption, only samples with more than 100 words were used. After collecting a large number of samples from the Common Crawl corpus, each sample's CWR was stored and mapped onto a distribution (Figure 2). The distribution is fairly normal, though there is a slight skew to the left. This implies that on a large number of samples, the distribution of CWR is fairly normal and resembles the distribution of IQs.

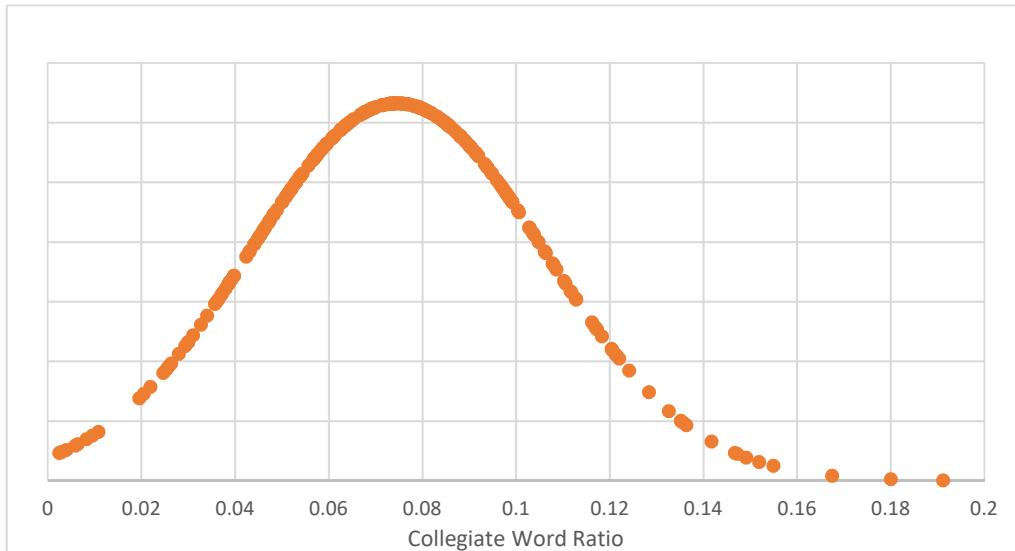


Figure 1: Collegiate Word Ratio Distribution

¹ The full list of words used for this project can be found at www.freewordlist.com.

² <https://aws.amazon.com/public-datasets/common-crawl/>

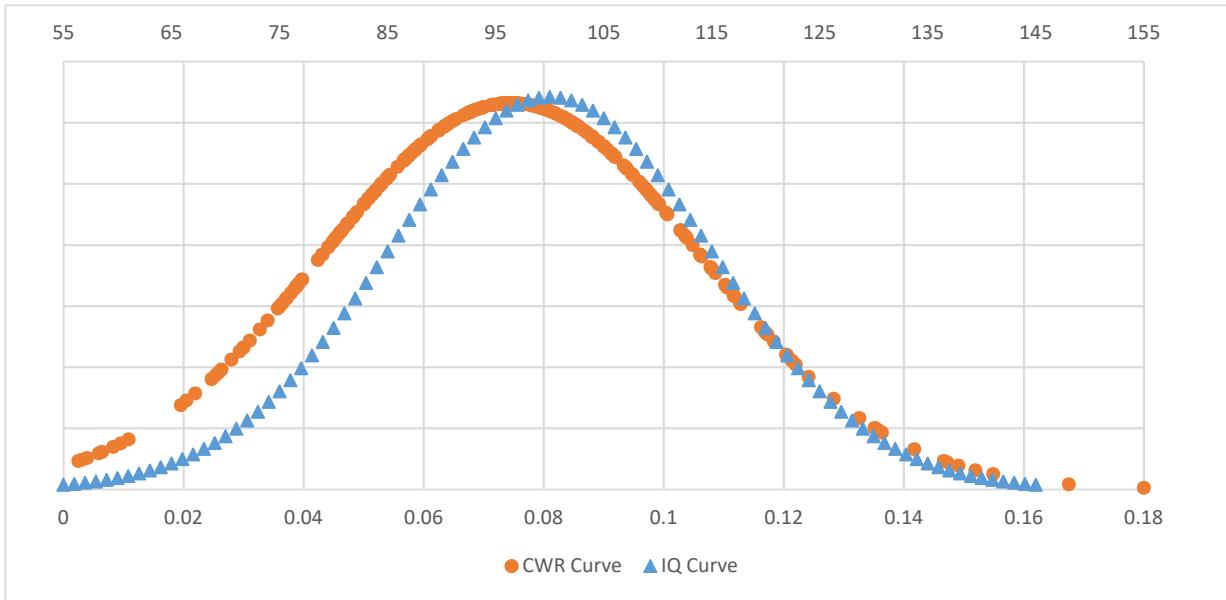


Figure 3: CWR VS IQ Distribution

3 Determining IQ from CWR

We have shown the distribution of CWR is fairly normal, and now we will demonstrate the process of using CWR to calculate an individual's IQ. A graph showing these two distributions overlaid is located below (Figure 3).

The IQ curve shown is the ideal expected IQ distribution. It is perfectly normal with a mean of 100. The CWR distribution, though skewed slightly left, is mapped very closely to the IQ distribution for the second and third positive standard deviation from the mean. For the purposes of this analysis, we will assume that this indicates the CWR in this area will map onto its corresponding IQ. This will result in a certain amount of error when calculating IQ from CWR. Nevertheless, the distributions are close enough that the process should give a good estimation of an individual's IQ.

To begin the process of transferring between the two curves, we need to know the standard deviation and

mean of both distributions. For the IQ curve, these are fixed values. The mean IQ value of all individuals is said to be 100 and the standard deviation of all IQ values is said to be 15. For our data set, the mean CWR is 0.074759005 and the standard deviation is 0.031552108.

Using these values and an individual data point's CWR, a corresponding IQ score can be calculated. Performing this calculation involves finding the z-score of the data point. This is done by the following:

$$\text{Z-Score} = (\text{CWR Data Point} - \text{CWR Mean}) / \text{CWR Standard Deviation}$$

Sample Word Length	Sample Collegiate Word Count	Sample CWR	Expected IQ	Measured IQ	% Error
752	94	0.1250	153	123.88	19.03
412	51	0.1238	130	123.31	5.15
136	22	0.1618	141	141.36	0.26
3279	433	0.1321	129	127.24	1.36

Table 1: Predicted VS Measured IQ

This z-score represents the number of standard deviations, positive or negative, that the data point is away from the mean. Since we know the standard deviation and mean of all IQ scores, the corresponding IQ can be calculated as follows:

$$\text{Corresponding IQ} = (\text{Z-Score} * \text{IQ Standard Deviation}) + \text{IQ Mean}$$

4 Testing IQ Estimation Software

Now that a sample of writing can be used to determine the IQ of an individual from their CWR, we must determine if the IQ is accurate. The process of doing this is straightforward, though difficult to accomplish. For it to be reliably said that CWR can be used to calculate an individual's IQ, we must find multiple individuals with a known IQ and access to writing that is their own. The pseudo-code for the software used to map the CWR of samples on to a corresponding IQ is shown in Figure 4.

Using social media contacts, we located several individuals willing to give their IQ and a sample of their

$$\text{Sample_Z_Score} = (\text{CWR_Sample} - \text{CWR_Mean}) / \text{CWR_Standard_Deviation}$$

$$\text{Sample_IQ} = (\text{Sample_Z_Score} * \text{IQ_Standard_Deviation}) + \text{IQ_Mean}$$

Figure 4: Calculating IQ Pseudo-code

writing for the purposes of testing our software. It should be noted that there is no external verification that these individuals gave an accurate IQ, but these samples are a good starting point for testing the reliability of this software. The data collected from these samples is shown in Table 1. Regardless of the large error in the first sample, the accuracy of the rest of the samples provide support for this approach for calculating IQ.

5 Conclusions and Future Work

Though our first sample produced a result with a moderate error, there still seems to be merit to looking further into this methodology. It should be noted that the samples used were approximately 2 standard deviations above the mean. Further sampling should include data on both ends of the curve. There may not ultimately be a cause effect relationship between intelligence level and vocabulary usage, but this research does indicate the two are correlated. The normality of the distribution of CWR may be significant in other applications, and should be noted regardless of the final merits of this approach to calculating intelligence.

This research paper is intended to be purely preliminary and simply introduce the concept and one possible implementation of using an individual's vocabulary to determine IQ levels. To further develop this research, the authors suggest a larger dataset be used to create a more accurate distribution. In addition, a more reliable dataset is necessary to test the accuracy of these methods. For the strongest possible results, self-reported IQ scores should not be used. Ideally, the next stage in research will include an IQ test along with a specific writing prompt on which to run our software. Lastly, there is likely merit in exploring the analysis of the other attributes of writing that are mentioned at the introduction to this piece. It is possible that one or all of these attributes may provide a better avenue for calculating an individual's intelligence level.

The ability to analyze the intelligence of individuals is a very useful tool. It has been shown in previous research that numerous factors influence whether an intellectually gifted child will ultimately lead a successful life (Tomlinson-Keasey, Little; 1990). Earlier identification of these children, through application of this research, has the potential to allow these children to be guided down a positive path that will lead to their personal success. In addition, this research could play a role in evaluating the abilities of persons currently prominent in the political and scientific realms. Nevertheless, further research must be done in this area of study before anything truly conclusive can be said.

6 References

- Ali, Nawaf; Hindi, Musa; Yampolskiy, Roman. (2011). Evaluation of authorship attribution software on a Chat bot corpus. Information, Communication and Automation Technologies (ICAT), 2011 XXIII International Symposium on, IEEE.
- Ali, Nawaf; Schaeffer, Derek; Yampolskiy, Roman. (2012). Linguistic Profiling and Behavioral Drift in Chat Bots. MAICS, 27-30.
- Ali, Nawaf; Yampolskiy, Roman. (2014). BLN-Gram-TF-ITF as a Language Independent Feature for Authorship Identification and Paragraph Similarity. 9th Cyber and Information Science research Conference, Oak Ridge, Tennessee.
- Bullerdieck, K. Kelly McK. (1985). Considerations in Defining the Gifted. <http://journals.sagepub.com/doi/abs/10.1177/10761758500800607>

- Franking, Holly. (1988). Stylometry: A statistical method for determining authorship, textual integrity, and chronology. University of Kansas, ProQuest Dissertations Publishing.
- Holmes, D. (1994). Authorship Attribution. Computers and the Humanities, 28(2), 87-106. <http://www.jstor.org/stable/30200315>
- Lawler, James. (1977). IQ: Biological Fact or Methodological Construct? *Science & Society*, vol. 41, no. 2, pp. 208–218. www.jstor.org/stable/40402014.
- Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist*, 42(2), 137-144.
- Tomlinson-Keasey, Carol; Little, Todd D. (1990). Predicting educational attainment, occupational achievement, intellectual skill, and personal adjustment among gifted men and women. *Journal of Educational Psychology*, vol. 82(3), pp. 442-455.
- Tweedie, F., Singh, S., & Holmes, D. (1996). Neural Network Applications in Stylometry: The "Federalist Papers" *Computers and the Humanities*, 30(1), 1-10. Retrieved from <http://www.jstor.org/stable/30204514>