5-2018

Automatic IQ estimation using stylometry methods.

Polina Shafran Abramov
University of Louisville

AUTOMATIC IQ ESTIMATION USING STYLOMETRY METHODS

By

Polina Shafran Abramov

B.A., Technion, Israel, 2004

A Thesis

Submitted to the Faculty of the

J.B. Speed School of Engineering of the University of Louisville

in Partial Fulfillment of the requirements

for the degree of

Master of Science in Computer Science

Department of Computer Engineering & Computer Science

University of Louisville

Louisville, Kentucky

May 2018

AUTOMATIC IQ ESTIMATION USING STYLOMETRY METHODS

By

Polina Shafran Abramov

B.A., Technion, Israel, 2004

A Thesis Approved On

April 24th, 2018

by the following Thesis Committee:

_____

Dr. Roman V. Yampolskiy, CECS Department, Thesis Director

_____

Dr. Ibrahim N. Imam, CECS Department

_____

Dr. Michael M. Losavio, Department of Criminal Justice

ACKNOWLEGMENTS

I would first like to express my sincere gratitude to my thesis advisor Dr. Roman Yampolsky. Dr. Yampolskiy provided me with all the guidance I needed, shared his ideas, at the same time giving me the freedom to make this work my own. He always found the time to respond to emails or meet in person to answer my questions. I am extremely grateful for his understanding and patience when I had to slow down.

I am thankful to Dr. Kantardzic for his invaluable advice throughout my years at Speed School and the time he took to meet with me. He always showed great interest in my work and took the time to discuss my studies and my future plans.

I would also like to thank all the people that offered their written samples and IQ scores for this research even though they didn't get utilized.

My sincere appreciation to Dr. Imam and Dr. Losavio for agreeing to be on my thesis committee.

Last, but not least, I must express my very profound gratitude to my husband who continuously supported and encouraged me throughout the years of my graduate studies and research, and to my family who always believe in me and support me in all my endeavors.

ABSTRACT

AUTOMATIC IQ ESTIMATION USING STYLOMETRY METHODS

Polina Shafran Abramov

Apil 24th, 2018

Stylometry is a study of text linguistic properties that brings together various field of research such as statistics, linguistics, computer science and more. Stylometry methods have been used for historic investigation, as forensic evidence and educational tool. This thesis presents a method to automatically estimate individual's IQ based on quality of writing and discusses challenges associated with it. The method utilizes various text features and NLP techniques to calculate metrics which are used to estimate individual's IQ. The results show a high degree of correlation between expected and estimated IQs in cases when IQ is within the average range. Obtaining good estimation for IQs on the high and low ends of the spectrum proves to be more challenging and this work offers several reasons for that. Over the years stylometry benefitted from wide exposure and interest among researches, however it appears that there aren't many studies that focus on using stylometry methods to estimate individual's intelligence. Perhaps this work presents the first in-depth attempt to do so.

Keywords: Stylometry, Artificial Intelligence, AI, IQ, Natural language Processing, NLP

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

1.1 Stylometry

Stylometry is a study of linguistic properties of the text which employs an analysis of various text features to study a document. Stylometry combines various fields of research such as statistics, linguistics, philology, psychology computer science and more. Perhaps the first instance of stylometry use can be attributed to Catholic priest Lorenzo Valla. In 1439, using philological arguments, he proved that the Donation of Constantine decree was in fact forged. Polish philosopher Wincenty Lutoslawski was the one who coined the term stylometry and defined the basics of it in Principes de stylométrie (1890).

Today, stylometry techniques are being applied in various areas such as academic research, disease detection, forensic evidence and more. In many cases stylometry requires processing of large amounts of data which was hard or even impossible to perform in the past. With development of computers, data analysis techniques, statistical tools and algorithms this task became much more feasible. The development of technology not only allowed for processing larger amounts of data, but also contributed to the ability to collect, store and grow data

corpora to be used in modern research. Today's stylometry efforts focus on extracting patterns, features and statistical information from text data whereas in the past its most common utilization was detecting and distinguishing the most interesting elements of the text.

1.2 Intelligence Quotient(IQ)

Before what in our days is known as IQ test was created, there were attempts to explore people's intelligence by observing their behaviors and analyzing their traits. The first test to measure intelligence was developed by Alfred Binet, Victor Henri and Théodore Simon in 1905. This test focused on verbal abilities. Eleven years later, in 1916, American psychologist Lewis Terman revised that test and created Stanford-Binet Intelligence Scales, which became the most popular IQ test in US for decades [1].

David Wechsler an American psychologist argued that there is a non-intellective factor when it comes to assessing intelligence and objected the single score of Binet scale. In 1939 he developed Wechsler-Bellevue Intelligence Test in which he divided the test into two main parts - verbal and performance (non-verbal) scales, each evaluated with different subtests [2]. Since then Standford-Binet test was revised to match Wechsler-Bellevue Test in several aspects, but the latter remains the most popular IQ test in US.

Generally, IQ score is calculated using the following formula:

$$IQ = \frac{MA}{A} 100$$

where MA is person's mental age score, obtained from an intelligence test, and A is person's actual age.

In modern IQ measures, the mean IQ score is defined as 100 and standard deviation of 15. Based on this, we can obtain normal distribution curve of IQ scores across entire population as shown in Figure 1 [3].

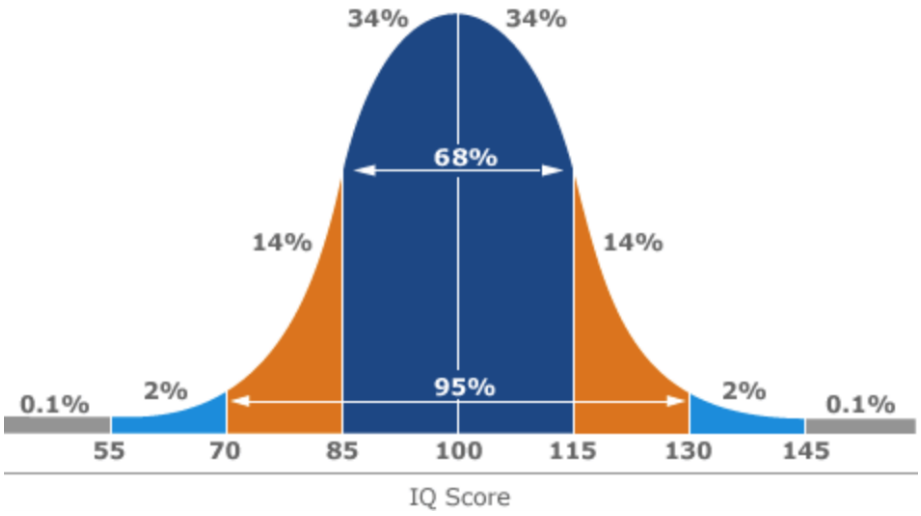

Figure 1. A normal distribution of IQ scores across entire population

LITERATURE SURVEY

Stylometry is a large topic that covers multiple areas of research. Some of those areas received more attention in the past years while others remain less explored. Modern development in computer science fields such as machine learning and natural language processing contributed to substantial advances in stylometry research.

2.1 Authorship Analysis

Authorship Attribution is one of stylometry categories that benefits from wide exposure and interest, partially due to the relative simplicity of the problem and data availability. In authorship attribution problem, we are given a list of possible authors and a document. The goal is to determine the most likely author. The most notable success of authorship attribution research dates to 1964 study of Mosteller and Wallace on the mystery of authorship of the Federalist Papers [4] . The satisfying results of the study gave validity to stylometry and initiated more studies in that area. Initial research focused on the attempt to define a set of features to determine writing style. That's when

measures such as sentence length, word length, word frequencies, character frequencies, and vocabulary richness were introduced. Later, the development of such areas as information retrieval, machine learning and natural language processing and the increased amount of digitally available data contributed to significant advances in authorship attribution research [5]. Despite a major success in this area of research authorship attribution remains a challenging problem.

A category of stylometry that received less attention is authorship verification. As opposed to authorship attribution, here we are given examples of the writing of a single author and are asked to determine if given texts were or were not written by this author. This problem proves to be significantly more difficult than authorship attribution problem. Moshe Koppel and Jonathan Schler explain this complexity as following: “If a text was written by Shakespeare or Marlowe, it would be sufficient to use their respective known writings, to construct a model distinguishing them, and to test the unknown text against the model. If, on the other hand, we need to determine if a text was written by Shakespeare or not, it is very difficult – if not impossible – to assemble an exhaustive, or even representative, sample of not-Shakespeare. “ [6]. The difference between authorship attribution and authorship verification problems is subtle but significant. While in authorship attribution we know that one of the candidate writers is the author, in authorship verification the candidate may or may not be the author. This distinction is the reason why authorship verification is a hard problem to solve.

Third main stylometry category deals with the attempt to identify author's personal traits, such as age, gender, origin, education etc. This category is referred to as Authorship Characterization or sometimes as Authorship Profiling. In authorship Characterizations researches try to use linguistic features and differences in how various groups of people speak or write to discover information about the author.

M. Koppel from Bar-Ilan University in Israel has done a significant amount of research in this area. He and his colleagues showed an approximately 80% success in identifying author's gender by combining stylometry and classification techniques on formal written text [7], ability to determine author's native language [8] and age [9].

S. Argamon et al explores even more interesting problem of trying to discover as much information as possible about the author by using content-based and style-based features [10]. Their research shows that an accurate choice of features and machine learning methods can help to find details about individual's demographics, background and personality.

2.2 Electronic Data

On one hand, vast amount of electronic texts available online provide a great and diverse data for future research. On the other hand, this data comes with its own challenges, such as shorter length and poor structure. Many

previously developed features do not work well with short text samples. When writing emails, posting on social media or sending a text message people tend to change their writing style by eliminating words, shorten sentences and avoid punctuation. Those eliminated items usually don't provide additional information for communication, but they are the ones that contain the information necessary for distinguishing between various writing styles. However, the situation with such data is not as desperate. Inability to rely on some known measures can be compensated by the metadata, such as email header information or attachments. In addition message structure or abbreviations used (e.g lol, btw, fwiw) can provide more clues on the author's identity [11].

Marcelo Luiz Brocardo et al. attempted to verify an authorship of emails of Enron's employees that were made public after the company bankruptcy. To overcome the issue of short messages length, the data was grouped by author to create a longer stream that is later divided into blocks. They suggest a model that generates a profile for each author based on the training block at the training stage and authorship checking at the verification stage. While the results are promising for certain block sizes, it is obvious that more research is required in this area [12].

2.3 Feature Selection

Feature selection is a difficult problem to tackle as there is no general agreement among stylometry researchers which feature should be used for which problem. It is common to select different set of features for different types of problems. For example, common words such as articles, pronouns and prepositions are usually excluded when performing topic based classification of a text, however those same words prove to be very useful for authorship attribution as they help to distinguish between various writing styles.

To make matters even more complex, in many cases the same features cannot be used on the same problem in different contexts, due to certain linguistic aspects not being shared by different languages, dialects and overall complexity of human language [13].

To our knowledge no large-scale research was performed to try and compare the effectiveness of various features across different problems. In fact, J. Rudman claims that most of attribution studies are done by a “one problem” practitioners making them focus on a specific problem without a lot of attention to the entire field [14]. Perhaps this can be justified, at least partially, by the large size and complexity of the stylometry research field and large variety of techniques and measures developed. For example, a computational tool Coh-Metrix [15][16][17][18][19] that offers metrics to calculate coherence of a text, contains 108 different indices. The tool was developed by Arthur C. Graesser and Danielle McNamara. “Coh-Metrix Measures Text Characteristics at Multiple

Levels of Language and Discourse“ [20] offers more in-depth information about Coh-metrix indices and architecture.

2.4 Applications

Aside from having many applications within the world of academia, stylometry has been utilized as educational tool, forensic evidence historic investigation and more.

Forensics investigators describe the usage of stylometry in helping to identify document authorships to solve crimes or address authorship disputes [21]. Various stylometry techniques can help solve crimes by identifying person’s origin, gender, education levels, age group and more. This can be achieved by examining spelling specifics, vocabulary differences and writing style. Despite success of some stylometry based evidences in court, J. Rudman talks about series of controversies and disagreements [14] that prevented the use of authorship studies in US courts. He also mentions Britain’s judicial system which accepts authorship attribution as a legitimate science. However, after one of its star expert witnesses had his method debunked on live television which presented, the judicial system was faced with a dilemma whether it made the right call by accepting such methods.

Stylometry also offers multiple applications in the world of education. Various measures exist to assess the readability of a text. The measure of readability (sometimes referred as text difficulty) can be vital for matching books with students based on their grade level [22]. The need in text difficulty measure is acknowledged in the Common Core Standards as well. Teachers are referred to Lexile Framework [23], whose goal is to match the reader with the text of the appropriate level. Lexile Framework uses Lexile Measure that represents a student's level on a developmental scale of reading ability— and matches it with student's grade equivalent. Automatic Essay Scoring (AES) is another application of stylometric methods. Its goal is to help mitigate rising education costs and support accountability by imposing standards [24]. Even though it has been criticized for various reason, AES is already being used in some schools to grade student's essays.

Additional areas of applications include but not limited to help with national security matters and market and history research.

2.5 Stylometry and IQ Assessment

For quite some time stylometry has been used to assess one's development level for education purposes. However, per our investigation not many studies attempt to detect the IQ level of text's author. Despite wide availability of various text corpora, one of the biggest challenges for such

research is finding the training and testing data. Ideally, for such research one would require not only the text corpora but also authors' IQ scores.

One attempt to explore a correlation between the Quality of Writing (QoW) and the writer's IQ was made by Nawaf Ali [25]. However privacy laws prevented him from obtaining access to the data required for the research and forced to change the original direction of research and settle with a simplified plan. In his study Ali is able to classify texts based on QoW using such features as occurrence of rare words, vocabulary richness, word's length and more. His results showed 99.8% accuracy when classifying texts of two highly distinct groups (Scientific Writing Samples vs School Students Writing) but proved more challenging when the borderline between intelligence groups was thinner, e.g. 4th-5th graders vs middle school students. A preliminary research "Automated IQ Estimation from Writing Samples" (A. Hendrix, R. Yampolskiy) [26] introduces the idea of correlation between the vocabulary used in a written sample and the writer's IQ. This research shows the existence of such correlation and urges further research on the subject.

In "The Other IQ" [27] Dean Keith Simonton talks about "historiometry" – a discipline in which the IQ assessment may be performed on participants that are long deceased, by applying quantitative analysis on historical data such as person's biography profiles, letters and political speeches.

This research might become the first and initial deep dive into the subject of stylometry based IQ assessment.

METHOD

Personal IQ information has privacy laws associated with it making it hard to gather real data for a research such as this one. Here, we are making an attempt to work arounds these limitations by proposing a hypothesis. Our method utilizes the bell curve distribution of IQ scores as shown in Figure 1. We are going to compute stylometric features on the training set and plot their normal distribution. The proposed hypothesis is that if the normal distribution of the computed feature matches the IQ scores distribution, then we can use the IQ curve to estimate author's IQ.

The analysis of our sample texts is performed using our proprietary python scripts and Coh-Metrix – a computational tool that produces indices based on various linguistic features of a text. We use the tools to calculate feature based indices that are then used to assess text samples. When it comes to Coh-Metrix, out of more than hundred indices of cohesion, language and readability that the tool generates we chose three that we believe represent the goal of this research. An additional fourth index is calculated using our own python script that utilizes NLTK library [28]. For pre- and post-processing of data several additional python scripts were implemented.

3.1 Features

In order to find a correlation between person's writing ability and IQ, we need to find a way to assess the quality of the written sample. A common way of doing it in stylometry is choosing several relevant text features and explore them. Our feature selection process relied on three aspects – previous research, experimentation and relevance. In her research on Linguistic Features of Writing Quality [29] Danielle McNamara et al. concluded that lexical features such as number of sentences, number of paragraphs, number of words per sentence and number of sentences per paragraph was not showing significant difference for high and low proficiency essays, hence those features were discarded. On the other hand, features such as lexical diversity and vocabulary proficiency showed correlation with individual's abilities. Multiple experiments were performed on more than 100 indices calculated by both our scripts and Coh-Metrix tool. Results that didn't show sufficient match between index's and IQ score's normal distributions were discarded. Lastly, multiple IQ test questions were explored and used as the guidance in selected appropriate features.

3.2. Selected Features

1. Lexical Aptitude Ration (LAR)

   For this feature, we utilize a list of words (denoted as D) that is used by SAT for evaluation of vocabulary proficiency. The goal is to identify

whether the author used any of those words in the text sample. Then given a text sample of length N, the formula for LAR is as follows:

$$LAR = \left\{ \frac{CountDistinct(W)}{N}, W \in D \right\}$$

2. Lexical Diversity (LDMTLD) is a measure of unique words used in the text. The simplest way to measure lexical diversity is to use type- type-token ratio (TTR) (Templin, 1957) that is defined as the number of unique words (called types) divided by the overall number of words in text (tokens). This measure, however, shows high sensitivity to text length. To reduce discrepancies caused by different lengths of text samples, we are going to use MTLD measure for Lexical Diversity, that was developed specifically to reduce the effect of text length. MTLD is calculated as the mean length of sequential word strings in a text that maintain a given TTR value [30].

3. Syntactic Complexity(SYNNP) measures the syntactic structure of the sentence. The sentence is considered less complex when, for instance, it has fewer verbs before the main verb of the main clause, when it is shorter or when it follows the simple syntactic pattern of actor-action-object. For this measure, we use Coh-Metrix SYNNP index which measures the mean number of modifiers per noun-phrase. A modifier is an optional element in a sentence and is said to modify (change the meaning of) another element in the structure, on which it is dependent. This is a good measure of working memory load.

4. Meaningfulness(WRDMEAc) feature is based on the meaningfulness ratings corpus developed by Toglia and Battig [31] that provides ratings for 2627 words. As Coh-Metrix description states “Words with higher meaningfulness scores are highly associated with other words (e.g., people), whereas a low meaningfulness score indicates that the word is weakly associated with other words.” [19] We use Coh-Metrix WRDMEAc index that calculates meaningfulness rating for content words only.

3.3 Data

3.3.1 SAT Vocabulary

A list of 5000 words for SAT preparation [32] is used to identify words for LAR feature.

3.3.2 Training Set

For training set we used Open American National Corpus (OANC) [33] that consists of texts of American English produced since 1990. The corpus includes both spoken and written text samples with written samples including technical articles, grant proposals, letters, essays and more. Only written texts are used in this research. The corpus has been preprocessed to exclude samples that are poorly written or constructed.

3.3.3 Test Set

Ideally, the test set would consist of text samples and the IQs of their authors. However, finding such set is a very hard task. This data is not publically available and not many people would willingly share it, especially if their IQ is relatively low. There are several people in the world with known IQ scores, for example, world renowned theoretical physicist Stephen Hawkings (IQ 160) and an American columnist and a writer Marylyn Von Savant (IQ 190). However, those are mostly people with an extraordinary high IQs which doesn't make for a balanced test dataset.

Selecting a text sample for these people would also be challenging as the goal and the target audience of these texts can vary, thus creating very incoherent data set. For example, if this is a scientific paper written for the audience of scientists, the choice of language and the structure of the text will take that into an account. In such texts, we can expect frequent appearance of field-specific terminology that is not as common outside the academia world, formulas and overall structure that is specific to scientific articles. On the other hand, if this same author were to write an article to be understood by the general public, chances are that the author would chose a simplified way to express ideas in "layman's terms". This creates a potential of constructing a non-homogeneous dataset that is hard to evaluate and compare.

To partially solve this issue, we used publicly available GRE sample essays as our test set [34][35]. There are several benefits in using these samples:

1. The samples are written on a given subject with the expectation for them to be evaluated and graded, hence offer a more homogeneous dataset.
2. The samples are written with the expectation to be evaluated and graded hence we can assume that the writer “did their best” when writing the text.
3. The samples are written by a single person and didn’t undergo any editing process.
4. Each text sample has been evaluated and analyzed by a human and given a score. The score can be used as an IQ estimation and mapped to an expected IQ.

Table 1

The interpretation of IQ Scores

IQ	Intelligence Level
> 130	Very Gifted
121 - 130	Gifted
111 - 120	Above Average
90 - 110	Average
80 - 89	Below Average
70 - 79	Cognitevely Impared

GRE scores for written samples go from 1 to 6 and are not as granular as IQ score. For this reason, each score is mapped to the range of IQ scores. Note that score 0 is also valid for GRE writing test, however for the purpose of this research we are discarding this score as it would indicate an empty text. In order to map GRE scores to IQ scores we use a chart that interprets the meaning of IQ scores shown in Table 1. The chart is based on Resing and Blok [36].

Table 2

Mapping of GRE writing samples scores to IQ score ranges

GRE Score	1	2	3	4	5	6
IQ range	70-79	80-89	90-110	111-120	121-130	131-160

GRE test is geared towards graduate students which are unlikely to have an IQ that is below average, hence mapping lower GRE grades to IQ ranges between 70 and 89 requires an additional explanation. A close examination of GRE samples that received lower scores shows that those are cases where an examinee either ran out of time or appeared as non-native speaker. Even though most likely those are not individuals with low IQs, their text samples can serve as an estimation for low-IQ samples. Following above logic, the mapping of GRE scores to IQ score ranges looks as shown in Table 2.

This research doesn't attempt to claim that there is a reliable way to convert GRE scores to IQ scores. We are aware that these two tests are different and there is no known correlation between GRE and IQ score. We are using only the samples from the Analytical Writing portion of the test to construct a homogenous set of written essays and simulate IQ scores. Our final test set contains twelve GRE text samples - two samples for each GRE score.

3.4. Process

3.4.1 Training

1. Preprocess OANC dataset.
2. Compute LAR, LDMTLD, SYNNP and WRDMEAc features.
3. Normalize computed features to match IQ range (40 - 160) and plot them as a normal distribution overlaid with the known IQ distribution curve. The first goal at this stage is to see how close the obtained distribution of text grades overlays with the IQ distribution curve.
4. Collect coefficients used in step 3 transformations. These coefficients are going to be used to transform test set results.

3.4.2 Testing

1. Compute LAR, LDMTLD, SYNNP and WRDMEAc features.
2. Use coefficients from Training step 3 to transform the indices of the testing set.

3. Evaluate the resulting score with respect to its proximity to the expected IQ range.

IMPLEMENTATION

4.1 OANC Dataset Preprocessing

OANC dataset contains large amount of text samples. Not all of them being relevant or useful for this research, hence certain degree of data preprocessing was required. The corpus includes text samples from various sources, including transcripts of spoken text. Due to the fact that this research focuses on written text, all spoken samples were removed from the training set.

The original corpora contained 6516 written text samples. During the analysis process, several samples that contained unreadable characters were discovered. Those samples could not be processed by automatic tools, hence were excluded.

Some of Coh-Metrix indices provide descriptive information regarding text sample, such as number of sentences, words and paragraphs. Out-of-norm values of those metrics can hint to poorly structured or poorly written text. For example, a text that contains only one sentence is either too short or completely lacks any punctuation, which would make it ineffective as part of training set. Coh-Metrix descriptive indices were examined to detect and remove such samples.

As a result of this preprocessing the remaining dataset that is being used as training set contains 5749 samples of written text.

4.2 Training Set Analysis

The calculation of the features on the training dataset was performed by our proprietary analytical program implemented in python using NLTK library and Coh-Metrix tool.

4.2.1 LAR Calculation

We use our own implementation to compute LAR. Our python script utilizes NLTK (Natural Language Toolkit) python suite that implements Natural Language Processing functionality.

The python script is reading the input text samples as raw text. In order to perform linguistic processing on it, first, it needs to be tokenized, i.e. converted to a structure of words and punctuations and then converted to NLTK text structure that provides wrapper for performing NLP operations.

```
import nltk
from nltk import word_tokenize

tokens = word_tokenize(raw)
text = nltk.Text(tokens)
lar = calculateLAR(text)
```

The calculation of this feature requires a predefined list of words that are considered proficient. We used a SAT preparation list of 5000 words, which was stemmed using NLTK Porter Stemmer [37]. This stemming is done in order to allow for a more flexible lookup in which we are looking for a word's stem rather than its exact appearance. For example, the SAT list includes the word "abridgment". Our goal is to detect all the cases in which this word appears in its various forms, such as "abridged" or "abridge". This becomes possible if instead of comparing the exact word we compare only its stem - "abridg".

```
from nltk.stem.porter import PorterStemmer

if __name__ == "__main__":
   f = open('vocabulary.txt', 'r')
   out = open('vocabulary_stem.txt', 'w')
   porter_stemmer = PorterStemmer()

  for line in f:
      sline = line.split(' ', 3)
      out.write(porter_stemmer.stem(sline[0]) + '\n')
```

Now that we have the list of stems, we can calculate the LAR index. Each word is stemmed before being looked up in the vocabulary. In order to improve performance, we skip stop words, such as "a", "an", "the", "and" as we can safely assume those words are not going to be on the list. There is an additional logic to account for cases when the same stem appears more than one time in the sample. We only count it once.

```
def calculateLAR(text):
   count = 0
   d = {}
   duplicates = {}

   with open("vocabulary_stem.txt") as f:
      for line in f:
         line = line.rstrip()
         if line not in d:
            d[line] = line
   for word in text:
      if (word not in stopwords.words('english')):
          porter_stemmer = PorterStemmer()
          stemmed_word = porter_stemmer.stem(word)
          if stemmed_word in d:
              #skip duplicates
              if stemmed_word in duplicates:
                  continue;
              duplicates[stemmed_word] = True;
             count+=1
  return count/len(text)
```

4.2.2 Features Calculated by Coh-Metrix

The three other features were calculated using Coh-Metrix tool. The resulting Coh-Metrix spreadsheet contains all 105 Coh-Metrix indices that were

calculated for each text sample. Out of those we select Lexical Diversity (LDMTLD), Syntactic Complexity (SYNNP) and Meaningfulness (WRDMEAc).

4.2.3 Features Transformation

Here the goal is to plot a normal distribution for each feature and to overlay it with the known IQ normal distribution. In order to do so, a linear transformation of a form *ax+b* is applied on each index to map its range to [40, 160] segment. This transformation is calculated separately for each index and performed using python script.

First we find the coefficients *a* and *b* by solving linear equation where *min_value* and *max_value* are the lowest and highest values of the given index.

```
def findCoefficients(min_value, max_value):
    return solve((40 - b - a*min_value, 160 - b - a*max_value), a, b)
```

Then, we apply the transformation on each value in the array of indices.

```
transformed_indices = list(map(lambda x:float(c[a])*x+float(c[b]), indices_arr))
```

One last thing to do is to move the transformed values so that their mean point aligns with the mean point of IQ standard deviation curve, which is equal to 100.

```
diff = 100 - np.mean(transformed_indices)
final_indices = map(lambda x:x+diff, transformed_indices)
```

4.2.4 Plotting the Data

After finding the a, b and diff coefficients and applying the transformation, the resulting index values are plotted along with IQ normal distribution. This allows us to assess the degree in which two curves align.

```
def drawPlot(indices):
    #plot indices
    indices = sorted(indices)
    mean = np.mean(indices)
    std_div = np.std(indices)
    fit = stats.norm.pdf(data_arr, mean, std_div)
    fig = plt.figure()

    pl.plot(indices,fit,'-o', color='yellow')
    pl.hist(indices, normed=True)

    #plot IQ normal distribution
    range = np.arange(lowest_iq, highest_iq, 0.019);
    pl.plot(range, stats.norm.pdf(range, 100, 15), color='red')
    pl.show()
    pl.close(fig)
```

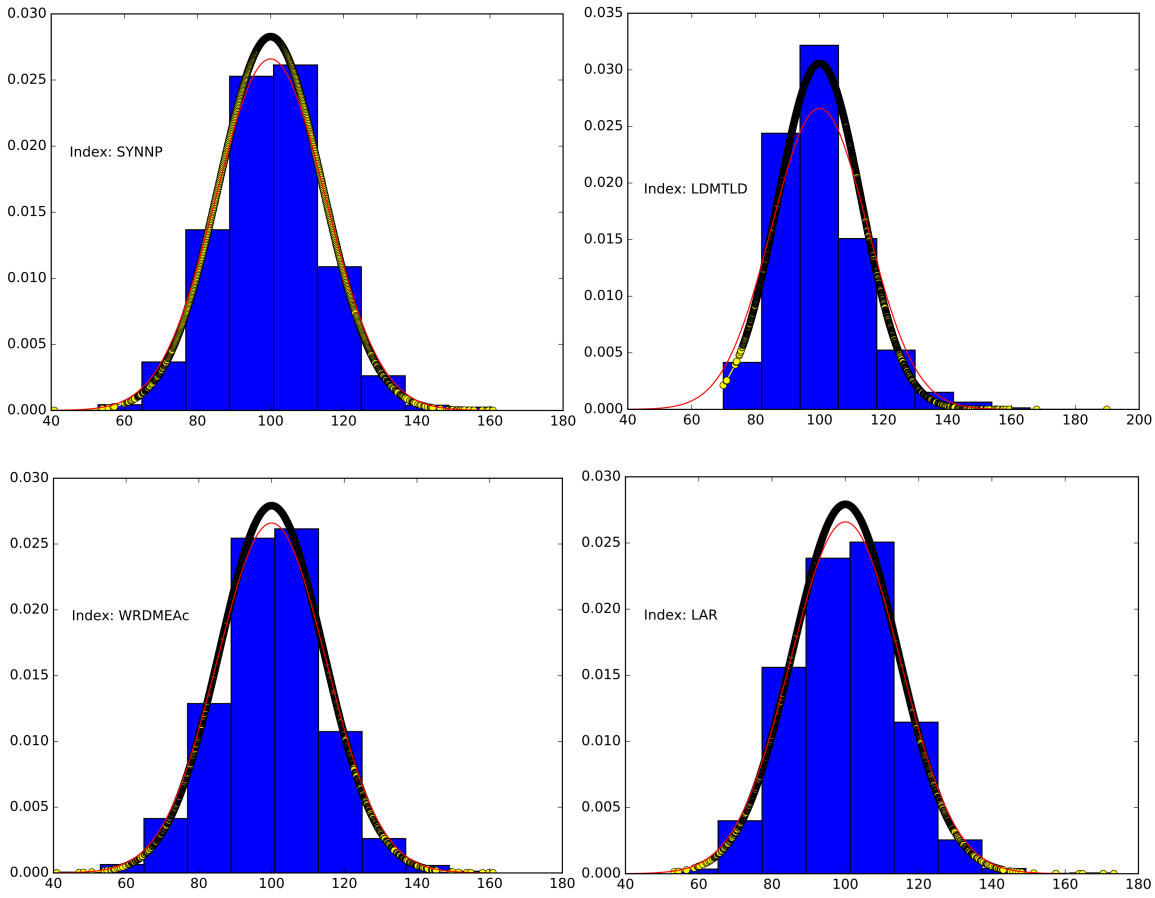

Figure 2: SYNNP, LDMTLD, WRDMEAc and LAR indices distribution (yellow) plotted with IQ normal distribution curve (red).

Figure 2 shows the resulting distribution for all 4 indices overlaid with the IQ Score normal distribution. Yellow curve represents the distribution of index values, while the red curve represents IQ bell curve.

4.3 Test Set Analysis

We are interested in calculating the same features for the samples from test set as the ones calculated for training set. As previously, the computation of LAR feature is performed by our proprietary analytical program implemented in

python using NLTK and LDMLTD, SYNNP and WRDMEAc features are computed by Coh-Metrix tool.

Having computed all four features for the test set, we used the coefficients that were calculated for the corresponding index from the training set in order to place the index value on the curve. This value is the Calculated IQ that we are going to compare for the Expected IQ. For example, for SYNNP index the calculation looks as follows:

$$Calculated\ IQ = a_{SYNNP} * SI + b_{SYNNP} + diff_{SYNNP} \quad (2)$$

Where SI denotes the test sample value of SYNNP feature and a_{SYNNP}, b_{SYNNP} and $diff_{SYNNP}$ are the coefficients calculated for SYNNP feature on the training set.

The final step of the process and consists of assessing the proximity of Calculated IQ to the Expected IQ. Since our Expected IQ is expressed as a range, we performed the assessment by calculating the error between the Calculated IQ and the high and low boundary of the Expected IQ range. If the Calculated IQ falls within Expected IQ range, then the error value is equal 0. Any value that has the error value less than 10% from either boundary is considered acceptable.

RESULTS

In this section we present the results of the analysis described in previous session. The analysis was performed on test set consisting of twelve GRE text samples. We experimented with various test sets before finally deciding to use GRE text samples. Using test samples from real people with known IQ scores yielded interesting results, however the main problems we ran into was lack of low or average IQ representation and the overall samples inconsistency. The texts differed so much in their structure and content that it was very difficult to perform a comparison between them. GRE text samples provided much more coherent dataset for our analysis, results of which is presented in Table 3 and Table 4. Table 3 shows calculated IQ Scores based on each one of the chosen features – SYNNP, LDMTLD, WRDMEAc and LAR. The left most column lists the Expected IQ that is compared with the Calculated IQ. Table 4 displays the results of this comparison by presenting the value of the error. The highlighted cells show all the results where the error is up to 10%.

The correlation between chosen features and IQ scores is visible in the obtained results even though not all of them fall within 10% margin. At least 60% of the results for each index estimate IQ level with up to 10% error with some indices showing particularly good results. For example, WRDMEAc feature

Table 3

Calculated IQ scores

Exp. IQ	Sample Name	SYNNP	LDMTLD	WRDMEAc	LAR
70-79	Sample 1	72.20	75.55	67.69	84.98
	Sample 2	104.98	76.11	123.90	98.12
80-89	Sample 3	131.28	76.96	72.54	95.93
	Sample 4	113.32	72.42	81.55	91.60
90-110	Sample 5	121.32	86.88	83.68	108.70
	Sample 6	114.83	84.59	93.51	125.56
111-120	Sample 7	108.01	88.98	104.92	121.93
	Sample 8	109.74	92.79	124.72	101.74
121-130	Sample 9	103.36	76.14	111.63	94.82
	Sample 10	119.37	117.02	104.47	135.67
131-160	Sample 11	118.08	95.65	121.67	89.20
	Sample 12	124.14	87.02	75.10	102.62

Table 4

Error values for Calculated IQ scores

Exp. IQ	Sample Name	SYNNP	LDMTLD	WRDMEAc	LAR
70-79	Sample 1	0.00	0.00	3.30	0.00
	Sample 2	32.89	0.00	56.84	24.20
80-89	Sample 3	47.50	3.80	9.32	7.79
	Sample 4	27.32	9.47	0.00	2.92
90-110	Sample 5	10.29	3.46	7.02	1.18
	Sample 6	4.39	6.01	0.00	14.14
111-120	Sample 7	2.69	19.83	5.48	0.00
	Sample 8	1.13	16.41	3.94	8.34
121-130	Sample 9	14.58	37.08	7.75	21.64
	Sample 10	1.34	3.29	13.66	0.00
131-160	Sample 11	9.87	26.99	7.12	31.91
	Sample 12	5.24	33.57	42.67	21.67

provides good estimations on the author's IQ level in 75% of the cases. Notably, the results for samples that represent non-extreme IQ scores (90-120) show very good approximation with 3 out of 4 indices showing errors within 10% range and

the remaining fourth index falling within 20%. As we observe the more “out of normal” IQ scores, the correlation is still noticeable but error values increase. For Sample 3 and Sample 4 we still see three out of four features giving a very close guess, but the error on the remaining fourth SYNNP feature gets almost up to 50%.

5.1 Weaknesses

Analyzing the results unveiled several weaknesses in our method. It is important to note that most of those weaknesses are present in standard IQ test as well and are not specific to our method, however they become more evident when using an automated method that does not involve an assessment by human.

5.1.1 Sample Length

Calculating text based metrics requires that a text sample is long enough to be analyzed. There is no single number of words that would be perfect for all cases, but from our experiments the minimum length requirement at which metrics give sensible results is around 300 words per text. Sample 2, for example consists only of 2 sentences and contains under 50 words, which without a doubt contributes to the difficulty in properly assess some of the features. It is interesting to note, that LDMTLD index for this sample shows an error that is less

than 10%, which complies with the claim that this specific index was designed to not be dependent on the length of the texts.

5.1.2 Extreme IQ Scores

In cases when IQ score is very low or very high, our method can be hard to rely on. People with IQ lower than 70 are classified as people with mental disability and the expectation to obtain a text sample that can be analyzed using normal metrics might be unreasonable. Same with the opposite case – the higher IQ score gets, the harder it becomes to solely rely on features of the text. Standard IQ test suffer from similar deficiency. Table 5 and Table 6 display results of IQ estimations for several individuals who are known to have extremely high IQ scores – S. Hawkins [38], Marilyn vos Savant [39], Garth Zietsman [40] and Anonymous M (personal info omitted for privacy reasons). The results are quite unsatisfying with error values varying between 20% and 40%.

Table 5

Calculated IQ scores for High IQ individuals

Exp. IQ	Sample Name	SYNNP	LDMTLD	WRDMEAc	LAR
160 - 180	Anonymous M	107.90	88.64	93.27	112.37
	Garth Zietsman	100.98	101.26	94.91	99.12
	Marilyn vos Savant	110.72	101.39	117.42	98.08
	S. Hawkins	104.12	91.54	93.11	108.17

Table 6

Error values for Calculated IQ scores for High IQ individuals

Exp. IQ	Sample Name	SYNNP	LDMTLD	WRDMEAc	LAR
160-180	Anonymous M	32.56	44.60	41.71	29.77
	Garth Zietsman	36.89	36.71	40.68	38.05
	Marilyn vos Savant	30.80	36.63	26.61	38.70
	S. Hawkins	34.93	42.79	41.81	32.39

5.1.3 Dependence on Language

The method in its current design is geared towards native English speakers as the indices are calculated based on English grammar rules. Furthermore, LAR feature relies on list of SAT words which is designed and used in United States, making the LAR index specific to American English. To make this method work for another language, one would need to calculate the same indices for that language. This limitation is not unique to our method. Regular IQ test is also language dependent, at least its verbal part, and requires an assessment using one's native language. In addition, just like regular IQ test, our method will potentially discriminate against individuals who are not using their native language to write the text sample. This isn't because of an inherent issue in our method design, but rather due to the fact that non-native speakers have a disadvantage when it comes to proficiency in foreign language as opposed to their native speaking peers. This can result in a less sophisticated text sample and lower IQ estimation.

5.2 Conclusion and Future Work

This work presents one of the first attempts to use stylometry principles to estimate individual's IQ score. Results obtained using our method are very promising and can serve as a stepping stone for further research in this area. One of the main things that would help to move this work forward is obtaining or creating a dataset of text samples with corresponding IQ scores of their authors. To avoid privacy complication, such dataset can be fully anonymized as we are not interested in specific identities, but rather the correlation itself. Having such training dataset will potentially allow researchers to achieve more precise results.

Four specific features were used in this research, however there is a lot of other information that can be extracted from a text sample and used to improve the assessment. Coh-Metrix tool offers more than 100 different indices and it is worth exploring them and their correlation with author's intelligence as well. Perhaps the assessment of text length could be incorporated into the analyses of the text to account for the edge case where the sample is too short to rely on calculated indices values.

Additional step forward would be to find an efficient way to combine results of various features into a single number that would provide the final estimation. The process of combining multiple features into one will need to be intelligent enough to account for different situations. Our results show that some features provide better estimation than others in different circumstances, hence

one should consider granting a different level of importance to each feature. The This can be done by assigning weights to each feature and calculating weighted average. The weights might need to be dynamic and change based on context. There is a potential to employ machine learning techniques such as genetic algorithm or neural networks to find the appropriate weight values.

REFERENCES

[1] "Intelligence Quotent," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Intelligence_quotient.

[2] "David Wechsler," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/David_Wechsler.

[3] "Statistics How To." [Online]. Available: http://www.statisticshowto.com/normal-distribution-probability.

[4] F. Mosteller and D. L. Wallace, "Inference in an Authorship Problem," *J. Am. Stat. Assoc.*, vol. 58, no. 302, pp. 275–309, 1963.

[5] E. Stamatatos, *A Survey of Modern Authorship Attribution Methods*, vol. 60. 2009.

[6] M. Koppel and J. Schler SCHLERJ, "Authorship Verification as a One-Class Classification Problem," 2004.

[7] M. Koppel, "Automatically Categorizing Written Texts by Author Gender," *Lit. Linguist. Comput.*, vol. 17, no. 4, pp. 401–412, 2002.

[8] M. Koppel, J. Schler, and K. Zigdon, "Determining an Author ' s Native Language by Mining a Text for Errors," *Proc. Elev. ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 624–628, 2005.

[9] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of Age and Gender on Blogging," *Artif. Intell.*, vol. 86, pp. 199–205, 2006.

[10] S. Argamon, M. Koppel, J. Pennebaker, and J. Schler, "Automatically Profiling the Author of an Anonymous Text," *Commun. ACM*, vol. 52, no. 2, pp. 119–123, 2009.

[11] R. Goodman, M. Hahn, M. Marella, C. Ojar, and S. Westcott, "The Use of Stylometry for Email Author Identification : A Feasibility Study," *Pace Pacing Clin. Electrophysiol.*, pp. 1–7, 2007.

[12] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, "Authorship verification for short messages using stylometry," *2013 Int. Conf. Comput. Inf. Telecommun. Syst. CITS 2013*, vol. 2013, no. Cits, 2013.

[13] C. Ramyaa, K. Rasheed, and C. He, "Using Machine Learning Techniques for Stylometry," *Conf. Mach. Learn.*, no. Proceedings of International Conference on Machine Learning, 2004.

[14] J. Rudman, "The State of Authorship Attribution Studies: Some Problems and Solutions," *Comput. Hum.*, vol. 31, pp. 351–365, 1998.

[15] N. M. M. Dowell, A. C. Graesser, and Z. Cai, "Language and Discourse Analysis with Coh-Metrix: Applications from Educational Material to Learning Environments at Scale," *J. Learn. Anal.*, vol. 3, no. 3, pp. 72–95, 2016.

[16] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich, "Coh-Metrix: Providing Multilevel Analyses of Text Characteristics," *Educ. Res.*, vol. 40, no. 5, pp. 223–234, 2011.

[17] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, “Coh-Metrix: Analysis of text on cohesion and language,” *Behav. Res. Methods, Instruments, Comput.*, vol. 36, no. 2, pp. 193–202, 2004.
[18] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press, 2014.
[19] D. S. McNamara, M. M. Louwerse, Z. Cai, and A. Graesser, “Coh-Metrix Web Tool 3.0,” *2013*. [Online]. Available: http//:cohmetrix.com. [Accessed: 20-Jul-2009].
[20] A. C. Graesser, D. S. McNamara, Z. Cai, M. Conley, H. Li, and J. Pennebaker, “Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse,” *Elem. Sch. J.*, vol. 115, no. 2, pp. 210–229, 2014.
[21] P. Juola, “Measuring Style: Document Analysis and Forensic Stylometry.” Juola & Associates.
[22] A. Metcalf, “Instant Readability,” *Chronicle*, 2016.
[23] A. J. Stenner and M. Smith, “Lexile Framework,” *MetaMetrics*. [Online]. Available: https://lexile.com/.
[24] “Automated Essay Scoring,” *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Automated_essay_scoring. [Accessed: 20-Aug-2003].
[25] N. Ali, “Text stylometry for chat bot identification and intelligence estimation.,” 2014.
[26] A. Hendrix and R. Yampolskiy, “Automated IQ estimation from writing samples,” *28th Mod. Artif. Intell. Cogn. Sci. Conf. MAICS 2017*, pp. 3–7, 2017.
[27] D. K. Simonton, “The ‘Other IQ’: Historiometric Assessments of Intelligence and Related Constructs,” *Rev. Gen. Psychol.*, vol. 13, no. 4, pp. 315–326, 2009.
[28] “NLTK platform Version 3.” [Online]. Available: http://www.nltk.org.
[29] D. S. McNamara, S. a. Crossley, and P. M. McCarthy, “Linguistic Features of Writing Quality,” *Writ. Commun.*, vol. 27, no. 1, pp. 57–86, 2010.
[30] P. M. McCarthy and S. Jarvis, “MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment,” *Behav. Res. Methods*, vol. 42, no. 2, pp. 381–392, 2010.
[31] M. P. Toglia and W. F. Battig, “Handbook of semantic word norms.,” *Handbook of semantic word norms*. Lawrence Erlbaum, Oxford, England, p. vii, 152-vii, 152, 1978.
[32] “SAT Vocabulary Words,” 2014. [Online]. Available: http://www.freevocabulary.com.
[33] “American National Corpus Project.” [Online]. Available: http://www.anc.org.
[34] “ETC,” 2011. [Online]. Available: https://www.ets.org/s/gre/accessible/gre_practice_test_2_writing_responses_18_point.pdf.
[35] “Sample Essay Responses and Rater Commentary for the Argument Task.” [Online]. Available: https://www.ets.org/gre/revised_general/prepare/analytical_writing/argument/sample_responses.

[36] W. Resing and J. Blok, “The classification of intelligence scores. Proposal for an unambiguous system,” *Psychologist*, vol. 37, pp. 244–249, 2002.
[37] “NLTK Stem Package.” [Online]. Available: http://www.nltk.org/api/nltk.stem.html.
[38] S. Hawkins, “This is the most dangerous time for our planet,” 2016. [Online]. Available: https://www.theguardian.com/commentisfree/2016/dec/01/stephen-hawking-dangerous-time-planet-inequality.
[39] M. v. Savant, “Logical Fallacies.” [Online]. Available: http://marilynvossavant.com/logical-fallacies/.
[40] G. Zietsman, “Noesis -The Journal of the Mega Society,” 2010. [Online]. Available: http://www.megasociety.org/noesis/190.htm.

CURRICULUM VITA

NAME: Polina Shafran Abramov

ADDRESS: J.B. Speed School of Engineering
Duthie Center for Engineering p0abra01@louisville.edu
University of Louisville
Louisville, Kentucky 40292

EDUCATION & TRAINING:

Master of Science in Computer Science
University of Louisville

B.A., Mathematics
Technion, Israel 2000 – 2004

WORK EXPERIENCE:

Software engineer with 15 years of experience in developing products ranging from personally tailored special-case solutions up to large-scale corporate systems.

SKILLS

- Languages: C/C++, C# , Objective-C, Python, PHP, VB.Net, HTML, XML, CSS
- Software Technologies: OOP/OOD, .NET Framework, iOS Frameworks, Design Patterns, MVC, Networking, RT Embedded, Cross Platform development, Multi-Threaded programming
- Development Tools: Microsoft Visual Studio .NET, C++ Builder, WinDbg, Dependency Walker, Eclipse, XCode, Cocoa, Tornado, Workbench (Windriver), InstallShield
- Libraries and APIs: Win API, Boost C++ Libraries, STL, GnuTLS, SMlib , OpenNurbs, SolidEdge API, Spatial
- Databases: SQL, SQLite
- Operating Systems: Microsoft Windows, Unix, Linux, VxWorks, iOS
- Developed for following CAD systems: Pro/E, Solid Edge, Solid Works, Catia, Optitex, Rhino
- Networks: HTTP/HTTPS, TCP/IP, UDP, RTSP, RTS, TLS/SSL, Whireshark

- Digital Video: MJPEG, TMP4, H.264, ONVIF, Motion Detection, Video Analytics

CareEvolution
2015 - Current
Software Architect

Image Vault
2011-2015
Lead Software Engineer

Ksoft
2008-2011
Lead Software Engineer

Nokia Siemens Networks (ex Atrica)
Embedded Software Engineer
2006-2008

Parametric Technology Corporation
Software Engineer
2004 – 2006