Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/42/20230773

Clustering algorithms for determining the number of authors in Bible

Yijie He

Department of Mechanical Engineering (Minor in Computer Science), The University of Hong Kong, Hongkong, 999077, China

u3609450@connect.hku.hk

Abstract. Ancient literature, exemplified by texts such as the Bible, carries immense cultural and historical significance, serving as a repository of the ancestors' experiences, triumphs, tragedies, and beliefs. Yet, delving into these ancient texts presents formidable challenges due to their limited accessibility. Over time, numerous historical documents have either vanished or undergone alterations, whether due to natural calamities or deliberate human actions. These obstacles have stymied progress in the field of diplomatics, rendering it a sluggish and occasionally unreliable endeavor. Remarkably, within very words and sentences of these texts lies a treasure trove of information. Properly harnessed, they offer a rich source of records for the advancement of diplomatics. For instance, deep learning techniques hold the promise of uncovering the number of authors who contributed to texts like the Bible. This research employs well-established and straightforward deep learning models—K-means, Gaussian Mixture Model (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)—to categorize the diverse writing styles present in the Bible. This methodology can be extended to other ancient English texts with uncertain authorship. The study involves the extraction of text from various versions of the Bible, which is then transformed into strings for analysis using these models. By categorizing different writing styles based on their underlying principles, the analysis facilitates an estimation of the number of authors who contributed to the Bible. Furthermore, the ensuing discussion offers insights into the advantages and limitations of this research project, shedding light on how its methods and findings might impact individuals.

Keywords: Deep Learning, Clustering, Authorship, Ancient Text, Bible.

1. Introduction

Deep learning is computational models consist of many processing layers to learn various characters of different layers in given data. This transformative technology has reshaped the world across various domains [1] e.g. gesture recognition, autonomous driving and facial recognition. Deep learning plays an important role in different industries nowadays because of its ability of study and adapt [2]. Clustering is an important unsupervised deep learning because it can put each of the unlabeled data from input into a structure which fit their characters [3]. Clustering is difficult already because of different size, density and shape of the countless kinds of data, the problem of data in very high dimension or be filled with noise [4].

DOI: 10.54254/2755-2721/42/20230773

One of the applications of clustering is Natural Language Processing (NLP) where it enables models to extract quantitative information from words within text [5]. NLP excels in performing a range complex tasks such as translation, customer service, emotion analysis, and more, thanks to its fundamental concept of enabling computers to comprehend and generate natural language text. [6]. The Large Language Models (LLMs) is a frontier technology based on NLP - its distinctive learning approach has ushered in significant opportunities [6]. Classifying text into different categories is an inseparable part of NLP, the text will be more valuable after processing and analyzing [7].

Throughout thousands of years of human civilization, countless books have been written. While some of these books contain detailed records of their authors, editors, and translators, many do not, particularly those authored by ancient sages centuries ago. Finding out number and names of the books' authors from ancient time can let know more about history of human civilization which is precious spiritual wealth of mankind. The traditional way to define the authors of aged books lacks efficiency and accuracy. Before, AI technology hasn't be used much in studying ancient text. With the rapid development of NLP, finding writers of ancient text have more chances to grow. Enhancing understanding of the various books that have exerted influence on the civilization for centuries would unveil greater historical truths, unravel mysteries, and facilitate a more robust inheritance of people's civilization, e.g., finding out how many authors of The Bible is a big progress in religion study. However, the application of deep learning on it has not been investigated in Bible studying much before.

In this research, 3 deep learning models will be used to cluster the content of The Bible to find out the number of different writing styles they have in order to infer how many authors The Bible has different versions of the most widely used Bible in English will be used in this research. Upon completing the research, the study will present the number of authors associated with The Bible and assess whether employing deep learning techniques in the analysis of ancient texts is both advisable and sufficiently beneficial. There will also be discussion about the future development and use.

2. Methods

2.1. Dataset description

2.1.1. Introduction of the data set

In this research, the Bible PDF files downloaded from various websites are employed. The Bible is in different version, there are 4 versions used which are the American Standard Version (ASV), the Catholic Public Domain Version (CPDV), the Darby Bible translation (DBT) and the King James Version (KJV). These include the most commonly used Bible among English speaking Christians around the world which has high representativeness, capable for this research [8].

2.1.2. Preprocessing of the data set

In the research, the models are directly applied to raw data without prior tokenization for clustering. However, pre processing is essential before clustering. The initial step involves vectorizing the text to enable computer comprehension. The text contains more than 4 million characters, so dimension reduction is required before clustering. The methods used in this research are autoencoder, TF-IDF and PCA.

Vectorization is essential in NLP and its result can easily affect the final output the the system [9]. The role of vectorization is mapping a high dimension data into a low dimension form so the data can have more applications.

The concept of Autoencoder was first introduced in 1986, it is a similar version of the PCA, its purpose is to obtain the useful information to get a compressed representation from the given data in an unsupervised way [10]. After reducing the dimension of the given data, the computation take place later will be easier, cost less time and memory [11]. During training autoencoder, the process can either be taken place layer by layers or trained as a whole [10]. TF-IDF consist of 2 parts, TF and IDF

DOI: 10.54254/2755-2721/42/20230773

[12]. Term Frequency (TF) means the number of times of a specific word appear in the text, if the word is a common word, the high TF value is unable to tell that this word is a representative of the text; Inverse Document Frequency (IDF) is a supplement for the TF, it reduce the influence of frequently used words and mention the importance of the rare words so the result will be more accurate and representative [13].

2.2. Clustering

In this research, 3 different methods will be used in finding out the number of writing styles in The Bible which are K -means, GMM and DBSCAN.

2.2.1. K - means

K means is one of the most commonly used and famous clustering methods [14]. K - means offers numerous advantages when compared to other methods. These include, but are not limited to, its simplicity in structure, swift computational speed, its ability to reveal clustering characteristics, and its consistent performance, even with large volumes of data, its basic logic is finding certain distance to on behalf of a sign for similarity at first, then decide a criterion function to define the quality of clustering, the center of clustering will be given next, finally the most reliable result will be given after recursion [15].

2.2.2. Gaussian Mixture Model (GMM)

GMM can represent combination of linear parametric probability density function of the input, it is usually used in the situation where the database has different distribution [16]. The general idea of GMM is the combination of all of the result of input vectors times the weight value and the probability, then the mixture density of the nth model is got [17].

2.2.3. Density based spatial clustering of applications with noise (DBSCAN)

DBSCAN was created in the late 1990s [17]. The fundamental concept behind this traditional model involves clustering data into distinct segments according to their density, followed by a search within a specified radius to determine the number of points in the vicinity where density may surpass a predefined threshold. Ultimately, the model provides insights into whether these points fall into the categories of core, broder, or noise [4]. DBSCAN is very good at dealing with high dimensions and other influences appear in the text to reduce scatter.

3. Results and discussion

The dimension reduction significantly deducts the time to have the result of the code for this research. Figure 1, Figure 2, Figure 3, Figure 6, Figure 7, Figure 8 are all the output histograms. The result is the number of different writing styles The Bible has according to the different models.

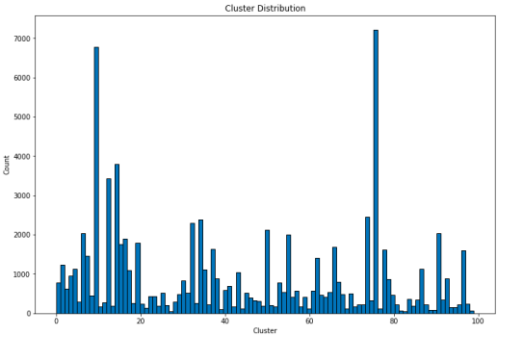

Figure 1. K-means clustering before autoencoder on ASV.

DOI: 10.54254/2755-2721/42/20230773

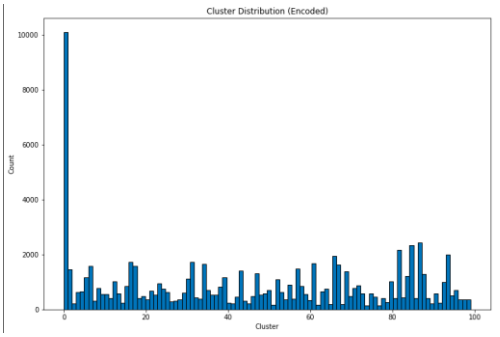

Figure 2. K-means clustering after autoencoder on ASV.

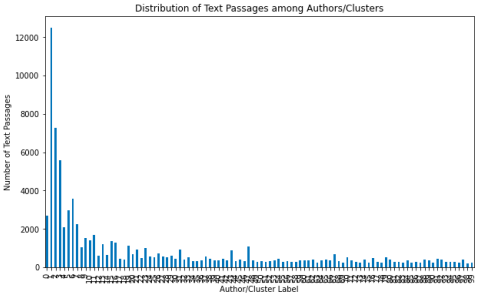

Figure 3. GMM before autoencoder on ASV.

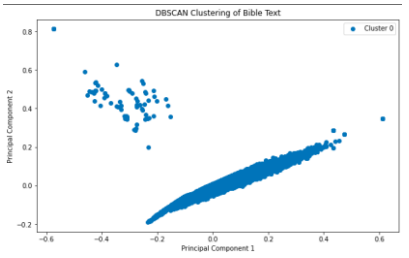

Figure 4. DBSCAN on ASV before dimension reduction.

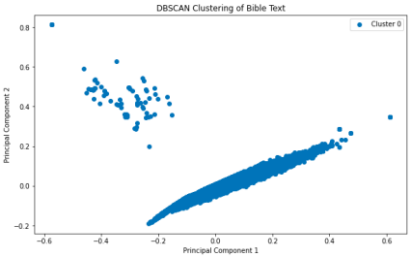

Figure 5. DBSCAN on ASV after dimension reduction.

DOI: 10.54254/2755-2721/42/20230773

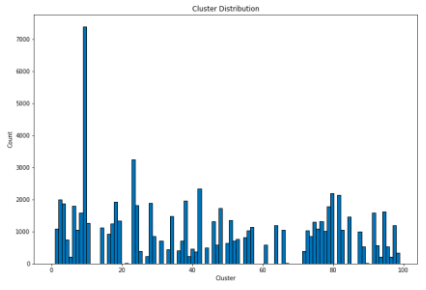

Figure 6. K-means clustering before autoencoder on KJV.

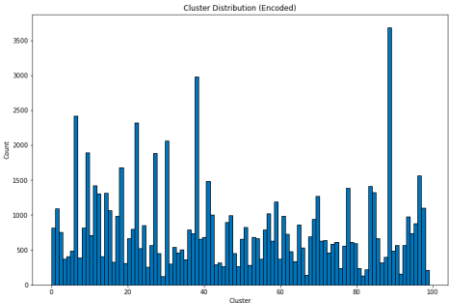

Figure 7. K-means clustering after autoencoder on ASV.

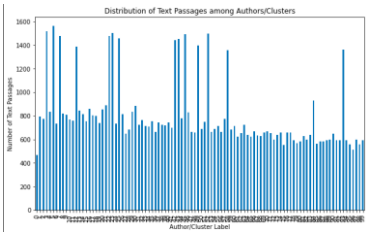

Figure 8. GMM before autoencoder on ASV.

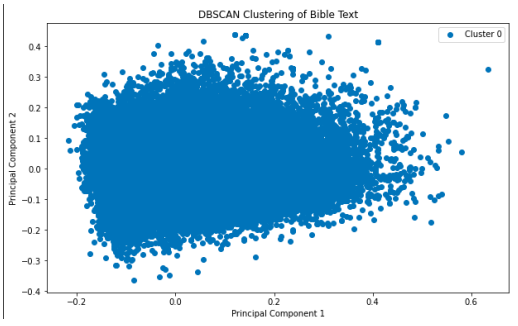

Figure 9. DBSCAN on KJV before dimension reduction.

DOI: 10.54254/2755-2721/42/20230773

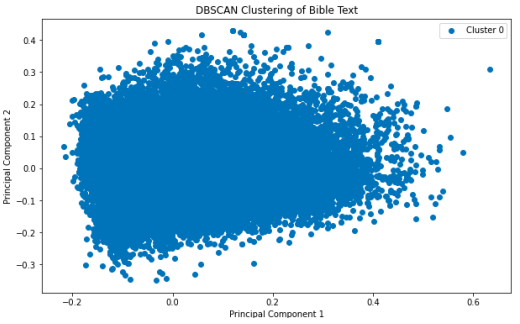

Figure 10. DBSCAN on KJV after dimension reduction.

Figure 1 to Figure 5 are the clustering results for the American Standard Bible (ASV) downloaded from https://biblehub.com/asv/genesis/1.htm; to be more specific, figure 6 is the result of K-means clustering, figure 2 is the result of K-means clustering after the vectors had been dimension reduction by autoencoder, figure 3 is the result of analysis by GMM model, figure 4 is the analysis result produced by DBSCAN, figure 5 is the result produced by DBSCAN after PCA do the dimension reduction.

From Figure 6 to Figure 10 are the clustering results for the American Standard Bible (ASV) downloaded from www.holybooks.com; to be more specific, figure 1 is the result of K-means clustering, figure 7 is the result of K-means clustering after the vectors had been dimension reduction by autoencoder, figure 8 is the result of analysis by GMM model, figure 9 is the analysis result produced by DBSCAN, figure 10 is the result produced by DBSCAN after PCA do the dimension reduction.

A more detailed explanation of the outcomes generated by the models are provided. When conducting the analysis using K-means and GMM models, a crucial parameter is required: the maximum anticipated number of clusters. In practical applications, these models consistently yield the number of distinct writing styles corresponding to the specified maximum number of clusters. To optimize the accuracy of the results, it is advisable to set this maximum number of clusters as high as system memory and GPU capacity allow. In the context of this research, the hardware setup can accommodate a maximum of 100 clusters. Consequently, the parameter has been configured accordingly, with a maximum value of 100.

In the ASV Bible analysis, before autoencoder's dimension reduction, there are 18 different writing styles have counts more than 1500, the highest count is 7224, 25 writing counts have more than 600 counts, the mean of all the writing styles' count is 874.2 and standard deviation is 1155.86, after autoencoder's dimension reduction, there are 13different writing styles have counts above 1500, the highest count is 10098, 47 writing styles have more than 600 counts, the mean all the writing styles' count is 874.2, the standard deviation is 1074.62; the GMM produced the result with maximum 10098, minimum 137, mean 874.2, median 578.5 and standard deviation 1074.62. The output produced by DBSCAN before and after the PCA dimension reduction have no big difference, there are 2 main parts for the writing styles to gather, the upper region is less dense than the lower one, the lower one has similar shape to linear function, only a few points are far away from the main part.

In the KJV Bible analysis, before autoencoder's dimension reduction, there are 15 different writing styles have counts more than 1500, the highest count is 7392, 48 writing counts have more than 600 counts, the mean of the top 48 writing styles' count is 1496.7 and variance is 3196793.0, the general situation of all the whole result is ;after autoencoder's dimension reduction, there are 9 different writing styles have counts above 1500, the highest count is 10389, 50 writing styles have more than 600 counts, the mean of the top 50 writing styles' count is 1298, the variance is 3496959; the GMM produced the result with 25 writing style with count more than 600, the highest is 12466, there are 10 different writing styles have counts more than 1500, the mean of the top 25 writing styles' count is 2290, the variance is 11750419. The output produced by DBSCAN before and after the PCA

DOI: 10.54254/2755-2721/42/20230773

dimension reduction have no big difference, both of them are a big semi elliptic consist of the dots present the writing styles with hundreds of dots near to it but still away from them.

From the result of the clustering by different models, the conclusion can be drawn which is the number of authors of the Bible is 36, the first 3 models analysis the ASV Bible support this result together; the lower division of gathered points in DBSCAN's result may tell the style of translation is similar and the authorship of the raw version is around 40, told by the points shown in upper division. The result of the first 3 models analyzing the KJV Bible also support this point of view; the semi elliptic may tell the style of translation while almost a hundred dots lie evenly across its right part so the Bible consist of countless different writing styles is proved. It is also summed up that the K-means, GMM and DBSCAN can cover each others' weakness in analyzing by their different characters. However, if there is more complex calculation required like setting the parameter higher like 200, 500 or even 3000 may tell more detailed result.

This research may build the groundwork to further or other research, solving the problems include but not limited to: The truth of John and Peter. Time and number of author for each chapter. Analyzing whether the some writing belongs to the same author, etc.

Conversely, this research also exhibits certain shortcomings that warrant enhancement. Firstly, the study employed only two versions of the Bible, despite the existence of numerous diverse Bible versions worldwide. A more comprehensive analysis involving a wider range of Bible versions could lead to more representative and robust conclusions. Subsequently, the versions of the Bible are in English only while there are so many versions of Bible in different language, the result may not represent Bible in other language. Additionally, the cluster is not detailed enough, setting the parameter higher may give more reliable result. Last but not least, other dimension reduction techniques should be used like PCA, t-SNE before clustering to avoid the influence brought by dimension reduction.

4. Conclusion

In this research, 3 deep learning models are used to detect how many different writings style the Bible has; 2 widely read versions of the Bible was used. The method for detection is clustering. Prior to the clustering process, dimension reduction techniques were applied to convert textual data into vectors. This step aimed to enhance computational efficiency and reduce memory usage. Subsequently, an analysis was conducted based on the models' outputs. The findings revealed that determining the precise number of authors responsible for the Bible's content is a challenging task. Instead, it was feasible to identify only the number of primary authors and editors involved in its creation. The result varies with different version of the Bible which is around 25 to 40. In the future, smaller division should be set to cluster more well and corporate with historians and literati to analysis the raw version of the Bible, analysis work for other ancient literature can also be done with the methods in this research.

References

[1] LeCun Y et al 2015 Deep learning. Nature 521, 436–444 https://doi.org/10.1038/nature14539

[2] Wang H et al 2020. Current status and application prospect of deep learning in geophysics. Progress in Geophysics (in Chinese), 35(20): 0642-0655, doi: 106038/pg2020CC0476

[3] Soni M T 2012 AN OVERVIEW ON CLUSTERING METHODS. IOSR Journal of Engineering, Vol. 2(4) pp: 719-725.

[4] Adriano M et al 2005 2005 Density-based clustering algorithms – DBSCAN and SNN. Pennsylvania State University. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=62272d87e82ffdec283c6da9d16f5065d7c44835

[5] Akira U 1996 Hierarchical Clustering of Words and Application to NLP Tasks. Retrieved Sep 16, 2023, from https://aclanthology.org/W96-0103.pdf

DOI: 10.54254/2755-2721/42/20230773

[6] Zhao T J et al 2023 Summary of Natural Language Processing [J/OL]. Journal of Xinjiang Normal University(Philosophy and Social Sciences). https://doi.org/10.14100/j.cnki.65-1039/g4.20230804.001
[7] Wi Z Y 2023Text classification and its uses based on Natural Language processing. Electronic Technology & Software Engineering (07),216-219. doi:CNKI:SUN:DZRU.0.2023-07-052.
[8] Sarah Z 2003 (n.d.) Christianity Today Retrieved Sep 17 2023, from https://www.christianitytoday.com/news/2014/march/most-popular-and-fastest-growing-bible-translation-niv-kjv.html
[9] Yi Y X 2018 The summary and analysis of text vectoring. Electronics World (22) 10-12 doi: 10.19353/j.cnki.dzsj.2018.22.003
[10] Bank D et al 2023 Autoencoders. In: Rokach, L., Maimon, O., Shmueli, E. (eds) Machine Learning for Data Science Handbook. Springer, Cham. https://doi.org/10.1007/978-3-031-24628-9_16
[11] Wu Z Y 2023 The use of text classification based on Natural Language processing. Electronic Technology & Software Engineering(07),216-219. doi:CNKI:SUN:DZRU.0.2023-07-052.
[12] Liu Z 2019 Application of TFIDF Algorithm in Article Recommendation System Computer knowledge and Technology (15)7 doi: 10.14004/j.cnki.ckt.2019.0959
[13] Jing Y J 2020 The principle and relization of new world recognition system based on TFIDF algorithm Information research (46)5
[14] Liu H et al 2018 Judgement method of vehicle lateral stability based on K means Clustering Analysis. Journal of Hunan University (Natural Science) 45(8) doi: 10.16339/j.cnki.hdxbzkb.2018.08.007
[15] Cao S Z et al 2010 The analysis of urban viaduct traffic flow based on K means algorithm. Traffic Technology (Applied technique version) 10 P261-264
[16] Hu Y J et al 2023 REsearch on garlic price prediction based on deep learning, Journal of Henan Institute of Science and Technology (Natural Science Edition) 51(3) doi: 10.3969/j.issn.2096-9473.2023.03.006
[17] Martin E et al 1996 A Density- Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA