Proceedings of IAM, V.12, N.1, 2023, pp.85-92
DOI: 10.30546/2225-0530.12.1.2023.85

PREDICTION OF AUTHORS OF THE TEXTS WRITTEN IN AZERBAIJANI BY THE NEAREST NEIGHBOR-BASED ALGORITHM

L.M. Hasanova

Baku State University, Baku, Azerbaijan
email: lamantry@gmail.com

Abstract The article analyzes the use of the nearest neighbor algorithm among the formal methods to identify particular authors of texts written in the Azerbaijani language and considers its application to the issue of recognizing specific authors. In the work under review, the roots of the words included in the texts of different authors are found, and the vectors according to the frequency of processing of these words are constructed according to a certain rule, and the nearest neighbor method is applied to them by treating these vectors as points in the Euclidean space.

Keywords: identification of authors, formal methods, nearest neighbor method, characteristic words, K-NN algorithm.

AMS Subject Classification: 91F20.

1. Introduction

Since ancient times, the issue of recognizing the author of the text has been an important problem. Thus, philologists, lawyers, historians, and many other experts are very interested in the topic of identifying the authors of written materials. In this regard, it is necessary to provide a number of suitable formal approaches for resolving the issue. It should be highlighted that approaches from the disciplines of image recognition, mathematical statistics, probability theory, neural networks, etc. are applied in the creation of formal methods [1-4]. It should be mentioned that [5] provides a general description of the primary formal approaches created for addressing the problem of attribution of writings and identifying their authors.

Recently, with the help of information technologies, the analysis of texts written in the Azerbaijani language, including the recognition of their authors, is being conducted [6-10]. In this field, the works of K. Aydazade and S.Talibov can be specially mentioned.

Formal approaches compare the features of texts calculated in a specific sequence. In this situation, the text is transformed into such a vector in which each of its constituents objectively represents some aspects of the text. The text is projected to a specific spot in n-dimensional space in this situation. In such a

formulation, such vectors may be used to identify each author, and this vector will be a vector obtained from the texts written by that author.

Following the definition of such vectors, we may discuss the proximity measure of two texts. This measure shall be the distance determined by any rule between the vectors corresponding to the texts. These vectors will, in the simplest instance, correspond to points in n-dimensional Decart space, the distance between which may be described as the usual Decart distance. Other choices, such as distance, can also be used. For instance, such a distance is regarded as the defining feature of a wide range of writings. Texts with a large distance between them can be attributed to different authors. Thus, it is necessary to calculate the corresponding parameters and determine the distance between the vectors made from these parameters, in order to compare the authors of any two texts. The vectors generated from the parameters corresponding to the author and the given text are compared in order to ascertain whether or not any text belongs to a particular author. To put it another way, two texts are still being compared. The standard text that belongs to the author is the first, and the controversial text whose authorship must be established is the second.

In addition to this, a vector of formal parameters may be built, allowing certain features of the writers to be determined. (for example, the level of education). It should be mentioned that some or all statistical properties of the text are used as text-characterizing parameters. These might include the quantity of particular words, punctuation marks, foreign terms, the number and length of sentences, the size of the vocabulary, average sentence length, and so on.It should be noted that one of the most widely used formal methods is the nearest neighbor method, or K-NN algorithm. This method is a classification method suggested by Cover and Hart [11].

Using the values of observations in a set of samples that are included in particular classes, the K-NN algorithm is used to determine to which class a new observation to be added to the sample belongs. This method is based on calculating the distances of each of the observations in the sample set to the subsequently determined observation value and selecting the class due to the number of observations with the smallest distance. For instance, suppose you wish to categorize a new object. At this moment, K classed elements are discovered to be the closest to it. If the majority of these items belong to the any class, the new element also belongs to that class. The nearest neighbor method is described in the image below (Fig. 1). The objects are separated into two groups, as seen in the image: blue squares and red triangles. The nearest neighbor method for determining which class the green circle object belongs to can be found as follows. If K=3, then 3 sample objects close to the object to be classified are found. As it is seen, two of these three items, or the majority of them, are classified as red triangles. As a result, the green unknown item will belong to the class of a red triangle. When K=5, three of the five items closest to the green sample belong to

the class of blue squares, implying that the green object does as well. This procedure can be done for every odd natural value of the parameter K.

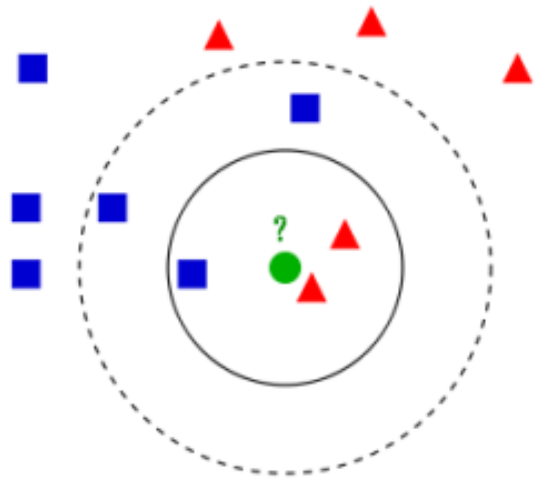


Figure 1. Example of K-NN nearest neighbor algorithm

İmportant parameters in the implementation of K-NN algorithm are distance, number of neighbors (k) and weight method.

Minkowski, Euclidean, Manhattan and Chebyshev distance concepts are used as distance measures. The Minkowski distance is a measurement method defined in Euclidean space. Classification is an extension of distance measurements such as Euclidean distance and Manhattan distance, which are commonly employed in clustering and data mining applications. The Minkowski distance between any two points $P = (x_1, x_2, \dots, x_n)$ and $Q = (y_1, y_2, \dots, y_n)$ is calculated according to formula

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} \quad (1)$$

The minkowski distance, expressed by a general formula, makes it possible to determine different distance measures for different values of equal p. Thus, in the Minkovsky distance formula, Euclidean distance is obtained when p=2, Manhattan distance when p=1, and Chebyshev distance when n→∞. In classification and clustering algorithms, the most common distance criterion is the Euclidean distance. This distance is calculated by formula

$$\left(\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\right) \quad (2)$$

Accordingly, the Manhattan distance is determined by the formula

$$\left(\sum_{i=1}^{n} |x_i - y_i|\right) \quad (3)$$

and the Chebyshev distance is determined by the formula

$$\lim_{p\to\infty}\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} = \max_{i=1,n} |x_i - y_i| \quad (4)$$

It should be noted that in classification and clustering algorithms like the K-NN algorithm, the primary distance criterion that is utilized to measure distance is the Euclidean distance.

Classification in the K-NN algorithm is based on the value of the number of neighbors (k) parameter. In the classification process, for k=1, only the nearest neighbor is assigned to the class, and as k approaches the number of samples (N), all information in the data set is taken into account and a selection is made according to the set number. In this article, author recognition system has been developed for the works of authors in the Azerbaijani language using the K-NN method, which is one of the methods of studying information.

In the work under review, the roots of the words included in the texts of different authors are found, and the vectors according to the frequency of processing of these words are constructed according to a certain rule, and the nearest neighbor method is applied to them by treating these vectors as points in the Euclidean space. It should be mentioned that there are many methods and algorithms related to the procedure of automatically finding the roots of words, in other words, the stemming operation [12-14].

2. The application of the nearest neighbor method

Now let's move on to the grating of attribute (signs) vectors corresponding to the authors. Suppose that the works we have belong to five authors. Let's assume that there are 10 works of each author, of which 5 works will be used as an example, and another 5 works will be used for testing. For each author, we join the initial 5 works that we took for instance, and dispose of a few additional components (focuses, commas, accentuation marks, numbers, and so on.) out of it.

To generate a character vector, the documents are first cleaned, which means that all punctuation marks and numerals in the documents are removed, and all letters are transformed to lowercase letters. First, the roots of the words in the combined document are identified, and the frequency of occurrence of these roots is determined and normalized. The frequency is calculated by taking the ratio of the number of the relevant word (more specifically, the word root) to the total number of words in the author's works.

After determining the frequency of words for each of the five authors, signs are chosen among them in the following sequence. Thus, for each author, n words (for example, n=10) are chosen such that their frequency of usage in this author's works is at least twice as high as their frequency in the works of other authors. To put it another way, the number of times these words are used by other authors should be less than half of what it is by the main author. This rule selects n=10 typical words for each author, which are then merged to form a 50-element vector. Thus, we will take as a basis the vector corresponding to 50 characteristic words, each element of which is the frequency of use of these characteristic words. After that, let's build 50-dimensional vectors corresponding to the selected characteristic

words for each work, both for samples (25 works) and for works that will be used for testing (50 works). The obtained vectors will be divided into 25 sample vectors (for works with known authors) and 25 vectors to be tested (for works with unknown authors), which may be thought of as points in a 50-dimensional Euclidean space. Now, let's apply the K-NN method to figure out who wrote the works that were tested. It should be noted that the K-NN technique mentioned above was presented for objects classified into two classes. In our situation, because there are five writers, the number of classes should be five. The objective is to determine which of these five classifications the unknown work belongs to.

3. Root-finding algorithm

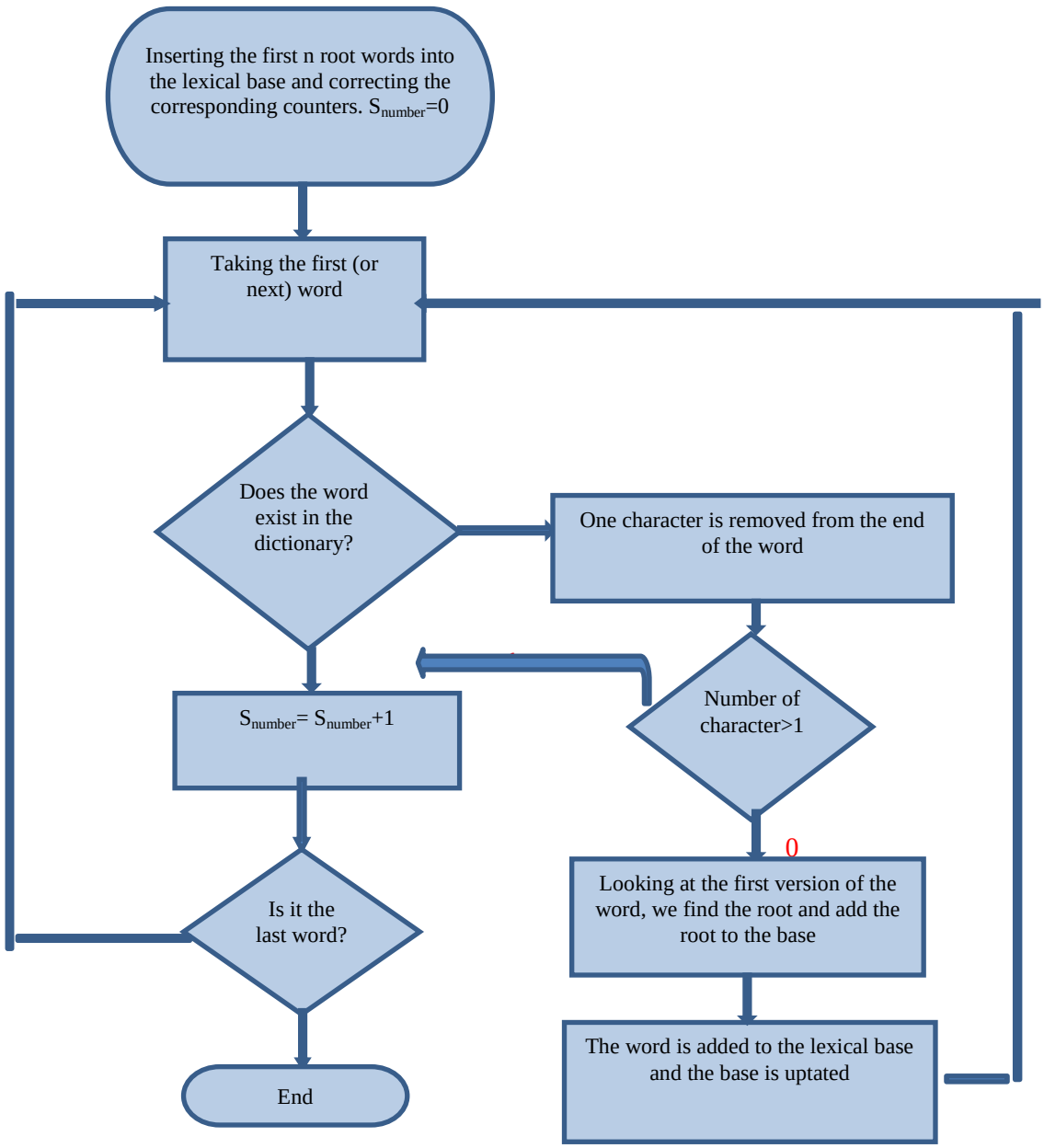
As previously stated, in order to identify the vector of signs, the frequencies of words in the work, or their roots, must be calculated.This can be accomplished by using an Azerbaijani lexical dictionary. Thus, by removing the suffixes from each word taken from the work in the correct sequence, the root of the term may be discovered and compared to the words in the lexical base. The value of the counter is then incremented by one unit whenever this word is met in the document by generating a counter for each word. After the document has been processed, the counters for each word will display the number of times that term appears in the document. We can calculate the frequency of word processing by dividing the value of these counters by the total number of words in the document (roots).

Now let's propose an algorithm for finding the root of the words in the work taken as an example and the number of corresponding words. In the event that there is no lexical dictionary database for the author, then we select n number of different words from the document related to that author, visually find their roots and create a lexical database for that author. The rule that follows determines the root of the subsequent word taken from the work using the algorithm that follows. To begin, we create a counter with the value 0 (zero) for each root in the lexical base. The following algorithm is then applied.

Algorithm

1. The next word is taken from the work in accordance with the sequence. (first word will be taken initially)
2. We compare this word with the words in the lexical base. If the word is found in the dictionary, we go to the next step, otherwise we go to the 5th step.
3. We increase the value of the counter corresponding to this word by one unit.
4. We proceed to the eighth and final step of the algorithm if all of the work's words have been analyzed. If not, we proceed to the initial step.
5. We take out the letter at the end of the word, and if the number of letters in the received word is more than 1, we go to the second step. Otherwise, we go to the next step.

6. We look at the initial version of the word and visually determine its root. The received new root is added to the lexical base and a comparing counter with an initial value of 1 (one) is created.
7. We return to step one following the addition of the new root to the lexical base.
8. End



4. Example

Let's see how this algorithm performs on a little piece of text.

Text

Bədəni kəndirlə sarınmış gənc, onlar çıxandan sonra o biri böyrü üstə çevrilib, otağa göz gəzdirdi. Yalnız bu zaman otağın bir küncündə büzüşüb oturmuş və heyrətlə ona baxan qadını gördü. Bu cavan və gözəl bir qadın idi. Gənc onu bir qədər maraqla süzüb, ağzına bulaşmış düşmən qanını təmizləmək üçün dalbadal bir neçə dəfə yerə tüpürdü.

First, we take n=5 and find the roots of 5 words from the text and enter them intothe lexical database.

Lexical base = {bədən, gənc, otaq, qadın, göz}

Let's take the values of $S_{bədən}$=0, $S_{gənc}$=0, S_{otaq}=0, $S_{qadın}$ =0, $S_{göz}$=0.

The word “Bədəni” is then extracted from the text as the first word. The term is searched in the database with upper case substituted by lower case. Since this word is not found, 1 character is deleted from the end of the word “bədəni” in the next step. Since the number of letters in the received word “bədən” is more than 1, this word is searched again in the database. Since it is found in the base, $S_{bədən}$=$S_{bədən}$+1 is executed and the value of the $S_{bədən}$counter is equal to 1. After that, the end of the cycle is checked. The next period begins as there are words left in the text.

As the next word, the word “kəndirlə” is taken. This word is searched in the database and not found. 1 character is removed from the end of the word consecutively and the received word “kəndirl” is searched in the database. If the word is not found, this process continues and the words “kəndir”, “kəndi”, “kənd”, “kən”, “kə” and "k" are searched in the database in the same way and are not found. Since the last word is a symbol, we look at the first version of the word “kəndirlə” and find the root “kəndir” and this word is added to the lexical base, that is, the lexical base = {bədən, gənc, otaq, qadın, göz, kəndir} and $S_{kəndir}$=1.

The process continues in this order. In one of the next steps, the words “gənc” and “Gənc” will be taken, and since this word is in the base, the operation $S_{gənc}$=$S_{gənc}$+1 will be performed twice, and as a result, the value of the $S_{gənc}$ counter will be equal to 2.

Finally, the word “tüpürdü” is taken as the last word in the text. Since this word is not found, this process continues according to the algorithm, and the words “tüpürd”, “tüpür”, “tüpü”, “tüp”, “tü” and “t” are searched in the database by deleting the last symbol and are not found. As the last word is a symbol, we look at the first version of the word “tüpürdü” and find the root “tüpür” and this word is added to the lexical base, that is, Lexical base= {bədən, gənc, otaq, qadın, göz, kəndir, ..., tüpür} and $S_{tüpür}$=1.

Thus, since the word “tüpürdü” is the last word of the text, the algorithm ends.

References

1. Aida-Zade K.R., Talibov S.G. Analysis of the effectiveness of the methods of recognition of authorship of texts in the Azerbaijani language, The 5th International Conference on Control and Optimization with Industrial Applications-COIA-2015, August, 2015, pp.27-29.

2. Aida-Zade K.R., Talibov S.G. Analysis of the methods for the authorship identification of the text in the azerbaijani language, Problems of information technology, N.1, 2017, pp.14–23.

3. Baayen, H., Van Halteren H., and Tweedie F. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, Literary and Linguistic Computing, V.11, N.3, 1996, pp.121-131.

4. Batura T.V. Formal methods for determining the authorship of texts, Vestnik NGU, V.10, N.4, 2012, pp.80-94.

5. Borisov L.A. Orlov Yu.N. Osminin K.P. Identification of the author of the text by the frequency distribution of letter combinations, Applied Informatics, V.26, N. 2, 2013, pp.95-108.

6. Cover T.M., Hart P.E. Nearest neighbor pattern classification, IEEE Transactions on Information Theory, V.13, N.1, 1967, pp.21–27.

7. Cover T.M., Hart P.E. Nearest neighbor pattern classification, IEEE Transactions on Information Theory, V.13, N.1, 1967, pp.21–27.

8. Gasimov S., Ibrahimov I. Analysis of sentences and words used in azerbaijani texts, The Second International Conference “Problems of Cybernetics and Informatics”, September 10-12, 2008, Baku, pp.117- 119.

9. Lovins J.B. Development of a stemming algorithm, Mechanical Translation and Computational Linguistics, V.11, N.1, 2, 1968, pp.22-31.

10. Mahmudov M. Computer linguistics, Baku: Science and education, 2013, 352 p.

11. Porter M.F. An Algorithm for Suffix Stripping, Program, V.14, N.3, 1980, pp.130-137.

12. Paice C.D. An evaluation method for stemming algorithms, SIGIR 94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1994, pp.42-50.

13. Romanov A.S. Technique and software package for identifying the author of an unknown text, Abstract of the thesis. dis. cand. tech. Sciences, Tomsk, 2010, 26 p.

14. Valiyeva K.A. Modern directions of computer linguistics, Information society problems, N.2, 2016, pp.98–107.