Central Asian Problems of Modern Science and Education

Volume 2020
Issue 3 *Central Asian Problems of Modern Science and Education 2020-3*

Article 8

11-19-2020

COSINE SIMILARITY AND ITS IMPLEMENTATION TO UZBEK LANGUAGE DATA

S.G. Matlatipov
PhD student, Applied mathematics and computer analysis department,National University of Uzbekistan, mr.sanatbek@gmail.com

COSINE SIMILARITY AND ITS IMPLEMENTATION TO UZBEK LANGUAGE DATA

Matlatipov Sanatbek G‘ayratovich
PhD student, Applied mathematics
and computer analysis department,
National University of Uzbekistan,
mr.sanatbek@gmail.com

Annotatsiya: Ushbu maqolada kosinus o'xshashlik va uning o'zbek tili matnlari o'xshashligini aniqlashga tatbiqi qaraladi. Shuningdek, maqolada kosinus o'xshashlikni aniqlaydigan dastur algoritmi bayon qilinadi. Muallif tamonidan taklif qilingan dasturning **ziyonet.uz** ta'lim portali matnlari kosinus o'xshashligini aniqlashga tatbiqi keltiriladi.

Kalit so‘zlar*: chastota vektori, normal chastota vektori, matn so'zlari teskari* chastotasi, *kosinus o'xshashlik, O‘zbek tili*

Аннотация: В этой работе рассматривается косинус подобия и его применение к определению сходства текстов на узбекском языке. Излагается алгоритм программы определяющего косинус подобия текстов. Приведём применение программы, предложенной авторами к текстам образовательного портала **ziyonet.uz;**

Ключевые слова: вектор частот, вектор нормальных частот, обратная частота текстовых слов, косинус подобие.

Abstract: In this article, it has been considered the cosine similarity and its application to search for similarities in the Uzbek language texts. The algorithm of cosine similarity has been used to determine similarity of Uzbek texts. We give the application of the program, proposed by the authors, to the texts of the educational

portal **ziyonet.uz** dataset;

Key words: frequency vector, normal frequency vector, inverse document frequency, cosine similarity, uzbek language.

I. Introduction

Development of natural language processing technologies available for low-resource languages is an important goal to improve the access to technology in Uzbek communities of speakers. Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis. If two vectors known from the geometry are parallel, the cosine of the angle between them is 1. If the vectors are perpendicular, the cosine of the angle between them is 0. The application of this rule to determine the degree of similarity of texts is a crucial and is very important in detecting plagiarism, automatic finding of novelty of ideas in the text, word embeddings as well as in search engines. Implementation of this metric can be applied to any two texts (sentence, paragraph, or whole document). Cosine similarity for similarity measurement between document and user query should accommodate to the word's meaning. Cosine similarity however still can't handle semantic meaning of the text perfectly [1]. Moreover, as Uzbek language considered low-resource language and deeply agglutinative, the similarity of words(texts) is expected to be noisy in the semantic meaning. However, similarity for documents of texts can be used the Cosine Similarity algorithm to work out the similarity between two things. The Cosine Similarity procedure computes similarity between all pairs of items. It is a symmetrical algorithm, which means that the result from computing the similarity of Item A to Item B is the same as computing the similarity of Item B to Item A. We can therefore compute the score for each pair of nodes once. We don't compute the similarity of items to themselves.

II. Related Work

The Cosine similarity algorithm implementation mostly for Turkic languages has been done exclusively on Turkish.

Comparison of various distance metrics and to determine the most appropriate methods in order to detect similarities among textual documents written in Turkish was studied in [2].

In [3] demonstrated how to analyse Turkish written text using Big Data technologies such as Nutch, Spark and MongoDB. More specifically, the work resolves operating the vectorization algorithms (Tf-Idf and Word2Vec) regardless of the language facilitates the study over a text whatever the text is written in any language. In Word2Vec feature extraction model, the word distances measured with cosine similarity.

Several classification methods presented based on the Paragraph Vector model devised [4] in document Embedding Based Supervised Methods for Turkish Text Classification. These include k-Nearest Neighborhood classifier (k-NN), Support Vector Machines (SVM), Centroid Classifier (CC) that works on paragraph vectors of documents and a custom made method which uses pairwise cosine similarities between documents and class centroids as features in Doc2Vec space.

Studying Turkish hypernymy using two supervised learning approaches based on word embeddings for offset between words pairs and semantic projection to link the words [5]. Each word is represented by its contexts in the form of a high-dimensional sparse vector and cosine similarity function is used for distributional similarity. There, The vector offset based methods using simple algebraic operations and cosine similarity have been successfully applied to analogy questions.

We can also find many papers for Kazakh language such as Auto-abstracting of text resources and documents in the Kazakh language [6]. The authors applied Machine

learning and cosine similarity of the sentence are calculated for determination of the sentence similarity.

III. Research Methodology

if text **A** is the same as text **B**, their cosine similarity is **1**. Conversely, if no word in text **A** overlaps with any word in text **B**, their cosine similarity is **0**. Hence, It has been concluded that cosine similarity can be applied to determine the degree of similarity of texts. Let's take a closer look at the cosine similarity for the texts. Given text **A**, let **B** be required to determine its cosine similarity to the text. To do this, worked on the following algorithm steps.

1-step

1) For each **A, B** text, we divide the texts into words (terms). We determine the number of words for each text. We will then refer to these words as the number of words in the text, respectively;

2) For each word in text **A**, we determine the number of times it occurs in text **A** (word frequency). We create a vector whose coordinates are these numbers. For each word in **A**, we determine the number (word frequency) of how many times it is present in text B. If this word does not occur in the text **B**, the number of this word in **B** (word frequency) will be **0** (zero). The number of coordinates of the resulting vectors is of the type **A** text words. We call these vectors frequency vectors;

3) We construct new vectors by dividing the coordinates of the frequency vectors corresponding to the texts A, B by the number of words in the text, and we call these vectors normal frequency vectors. For texts A and B, let these vectors be $\overrightarrow{X_1}$ va $\overrightarrow{X_2}$, respectively;

2-step

The concept of Inverse Document Frequency[7] also applied for Uzbek words The purpose of this concept is to find the texts that are as (closest) as possible to the

given text. There are words that are almost insignificant in terms of the degree of similarity, as they are very common in the text (later we will call such words insignificant words) and, conversely, there are words that are rare in the text, but The degree of similarity is very important in terms (we will later call such words important words). We feel that increasing the frequency of important words and reducing the frequency of unimportant words brings us closer to the goal.

$IDF(x) - x$ let the text be the inverse frequency of the word. We calculate the $IDF(x)$ as follows:

$$IDF(x) = 1 + \ln(\frac{n}{n_x});$$

n - the number of all texts viewed. (N = 2 for the problem we are looking at);

$n_x - x$ the number of tables that contain the word (n_x is at least 1 for the problem we are looking at).

Calculation of the inverse frequency of the text for each word in the text **A** and construct a vector **I** whose coordinates consist of these numbers (where the number of coordinates of the vector is of the length A in the text). this vector **A** the inverse frequency vector of the text;

3-step

For texts **A** and **B**, we construct new vectors by multiplying each coordinate of the normal frequency vectors $\overrightarrow{X_1} and \overrightarrow{X_2}$ by the coordinate of the inverse frequency vector I, respectively, and denote these vectors as $\overrightarrow{V_1} and \overrightarrow{V_2}$;

4-step.

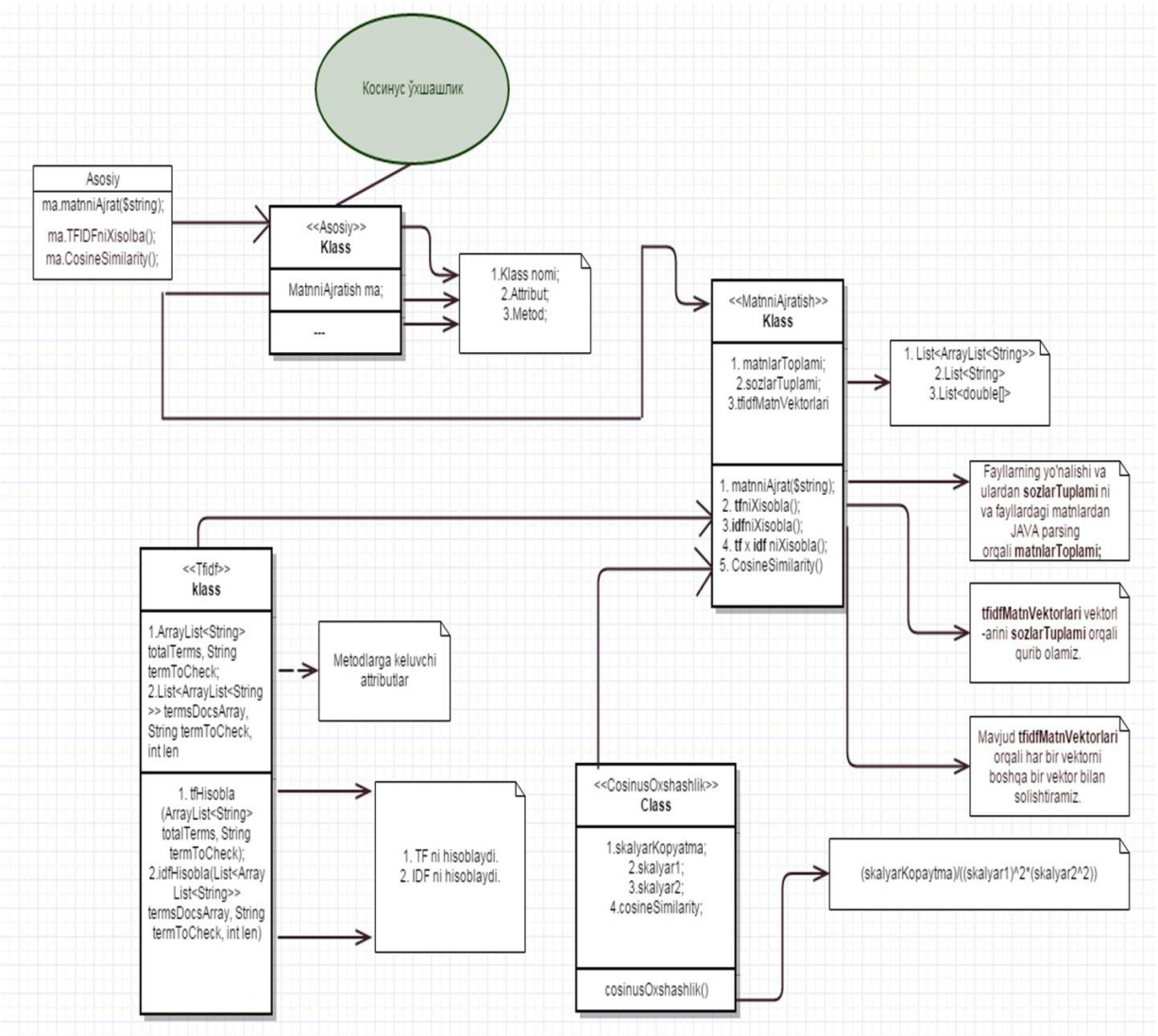
Let α be the angle between the vectors $\overrightarrow{V_1} and \overrightarrow{V_2}$. Then we calculate the following

$$\cos(\alpha) = \frac{\overrightarrow{V_1} * \overrightarrow{V_2}}{|\overrightarrow{V_1}| * |\overrightarrow{V_2}|};$$

Here, $|\overrightarrow{V_i}|$ - $\overrightarrow{V_i}$ length of the vector;

A program that detects the cosine similarity of Uzbek texts.

Based on the above algorithm, a program that determines the cosine similarity of texts is written in the Java programming language.[1] Below is the structure of the program



IV. Analysis and Results

It is usually important to determine if the work done is plagiarized[8] or not plagiarized when students write an essay, course work, graduate qualification work. this problem can be solved using the cosine similarity problem. In this section of the article we will see the application of the program based on the above algorithm to the Uzbek language texts.

[1] https://github.com/SanatbekMatlatipov/Cosine-similarity-for-Uzbek-texts

The following is the result of applying the cosine similarity to the texts of the educational portal **ziyonet.uz**

№	File names	Size
1.	1_Kollej talabalarini iqtisodiy savodxonligini oshirishda innovatsion texnologiyalardan foydalanish yo'llari**.doc**	(60-pages), 535KB
2.	2_Serfayz o`zbek dasturxoni**.doc**	(55-pages), 1476KB
3.	3_Birinchi sinf o'quvchilarida musiqiy savodxonlikni tarkib toptirish yo'llari**.doc**	(51-pages), 962KB
4.	4_Geografiya darslarida "Geografik maydonchadan" foydalanishning amaliy ahamiyati**.doc**	(77-pages), 2205KB
5.	5_Ta'lim jarayonida yoshlarning intellektual ijodiy qobiliyatlarini rivojlantirishning psixologik asoslari**.doc**	(56-pages), 217 KB
6.	6_Aqliy rivojlanishida nuqsoni bo'lgan bolalarni ta'lim jarayonida kasb-hunarga yo'naltirish omillari**.doc**	(54-pages), 240 KB
7.	7_Darsdan tashqari ta'lim-tarbiya jarayonida o'quvchilarda ekologik madaniyatni tarbiyalash**.doc**	(54-paages), 332 KB

The text files in the table (http://library.ziyonet.uz/uz) are given, The cosine similarity between the 1-text (*1_Kollej talabalarini iqtisodiy savodxonligini oshirishda innovatsion texnologiyalardan foydalanish yo'llari**.doc***) file itself and the rest of the files was determined. Using the program, we present the result table of their cosine similarity:

Comparable text files		**Cosine similarity**	**Performance time**
1- text file	1- text file	1.0	6.23 sec

1- text file	2- text file	0.5141095471924675	14.9 sec
1- text file	3 text file	0.41218812976955926	6.50 sec
1- text file	4- text file	0.45904909457705473	6.23 sec
1- text file	5- text file	0.572526390153752	3.44 sec
1- text file	6- text file	0.3612362022808788	2.89 sec
1- text file	7- text file	0.5724312304378543	3.86 sec

V. Conclusion

The article discusses the cosine similarity of Uzbek texts and its practical significance. Also, its statistics were given using the cosine similarity detection program in the texts of the educational portal **ziyonet.uz**

References

[1] Rahutomo, Faisal & Kitasuka, Teruaki & Aritsugi, Masayoshi. (2012). Semantic Cosine Similarity.

[2] Kaya Keleş, Mümine & Özel, Selma. (2017). Similarity detection between Turkish text documents with distance metrics. 316-321. 10.1109/UBMK.2017.8093399.

[3] Cakir, Ulas & Guldamlasioglu, Seren. (2016). Text Mining Analysis in Turkish Language Using Big Data Tools. 614-618. 10.1109/COMPSAC.2016.203.

[4] H. İ. Çelenli, S. T. Öztürk, G. Şahin, A. Gerek and M. C. Ganiz, "Document Embedding Based Supervised Methods for Turkish Text Classification," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, 2018, pp. 477-482, doi: 10.1109/UBMK.2018.8566326.

[5] Savaş Yıldırım and Tuğba Yıldız "Learning Turkish Hypernymy Using Word Embeddings", International Journal of Computational Intelligence Systems, 11/1, pp. 371-383, doi: https://doi.org/10.2991/ijcis.11.1.28

[6] ICEMIS'20: Proceedings of the 6th International Conference on Engineering & MIS 2020September 2020 Article No.: 96 Pages 1–5 https://doi.org/10.1145/3410352.3410832

[7] Robertson, Stephen. (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. Journal of Documentation - J DOC. 60. 503-520. 10.1108/00220410410560582.

[8] Habibulla Madatov, San'atbek Matlatipov. "Plagiat va uni fosh qilish dasturlari haqida". UrDU ILM-SARCHASHMALARI.2014-yil. 78-80 betlar.

[9]. Allan, J., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 314-321). ACM.