Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 53 – 62

Procedia
Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Text Classification Using a Novel Time Series Based Methodology

Zeev Volkovich[a] and Renata Avros[a]*

[a]*Ort Braude College of Engineering, Karmiel 21982, Israel*

Abstract

This paper discusses a novel time series methodology for writing process modeling, taking into account the dependency between sequentially written text parts. A series of consecutive sub-documents of a given document are represented via histograms of the appropriately chosen terms. To characterize the document overall style and its fluctuations, a new feature named the Mean Dependence is introduced. This similarity measure quantifies the association between a current sub-document and numerous earlier composed ones. So, such a collection of sub-documents is represented as a time series of the Mean Dependence development. The series change points naturally link to the style changes. Two possible approaches constructed within the general methodology are discussed. The first one intended to study media sources, is constructed to detect change points of media associated with social life transformations. Consequently, the homogeneous periods are detected using a new distance based on the Mean Dependence. The proposed methodology is applied to analysis of editorial texts published in the Egyptian "Al-Ahraam" and succeeds to indicate several important events connected to the "Arab Spring". The second approach, based on the strictly stationary model of time series, is applied to authorship verification. Numerical experiments demonstrate high ability of the proposed methods to recognize an authorship and to expose writing style evolution.

Keywords: Text Classification; Time Series Text Model; Authorship Verification

* Corresponding author. Tel.: +972-4-990-1994; fax +972-4-990-1852.
E-mail address: vlvolkov@braude.ac.il

doi:10.1016/j.procs.2016.08.095

1. Introduction

With the huge number of on-line documents abounding on the Internet, text classification has come to be one of the crucial methods for treatment and systematizing text data. Such tasks arise in the authorship recognition, automatic media content analysis, plagiarism detection and other areas. The main weakness of the methods used in these fields is that the association between texts is frequently evaluated without any regard to their developing process. Such as, for newspapers the similarity assessment of the issues is not accompanied by any connection to the early published ones. One of the common viewpoints of the human writing process views this process as composition of four key elements: planning, drafting, editing, and writing the final draft. Thus, it is natural to presume that dependency between sequential written text parts is remained at the almost uniform level if the text is composed by the same author, or in case of an official newspaper the social situation is relatively stable. This dynamical modeling of the writing process is the main advantage of the proposed method.

This paper discusses a novel time series methodology to the writing process modeling taking into account the dependency between sequentially written text parts. A series of consecutive sub-documents of a given document are represented via histograms of the appropriately chosen terms. To characterize the document overall style and its fluctuations, a new feature named the Mean Dependence is introduced. This similarity measure quantifies association between a current sub-document and numerous earlier composed ones. So, such a collection sub-documents is represented as a time series of the Mean Dependence development. The series change points naturally link to points of the style changes. Two possible approaches constructed within the general methodology are discussed. The first one intended to study of the media sources, is constructed to detected change points of media associated with the social life transformations. Consequently, the homogeneous periods are detected using a new distance based on the Mean Dependence. The proposed methodology is applied to analysis of editorial texts published in the Egyptian "Al-Ahraam" newspaper and successes to indicate several important events connected to the "Arab Spring". The second approach based on the strictly stationary model of time series is applied to the authorship verification. Numerical experiments demonstrate high ability of the proposed methods to recognize an authorship and to expose of a writing style evolution.

The rest of the paper is organized in the following way. A short review of related works is presented in Section 2. The proposed general methodology is stated in Section 3. Section 3 demonstrates an application of the suggested methodology to analysis of editorial texts published in the Egyptian "Al-Ahraam" newspaper for the periods containing some important events in the political life of the appropriate society, particularly, the "Arab Spring". A modification of the methodology based on a strictly stationary model of the Mean Dependency development is given in Section 4 together with an approach to the author verification problem and the appropriate numerical experiments.

2. Related Works

The field of authorship attribution originates from stylometry, analyzing texts for evidence of authenticity, authorial identity, and other questions. The task has a long history and an overview of different methods in chronological order is given in surveys [1,2]. One of the essential parts in quantitative authorship attribution algorithms is the distance measure quantifying the similarity between the texts. Burrow's Delta[3] is the most recognized measure of stylistic difference. Since its first appearance in 2002, various modifications have been proposed. The performance tests for Delta and its variants are provided in [4]. The compression-based measures, like application independent Normalized Compression Distance, are successfully applied to the text clustering. The comparison of numerous compression models for authorship attribution is performed in [5]. Despite their universality these measures are computationally expensive and therefore difficult to use in practice. To reduce the computational complexity in [6] a new model is developed.

Character *N*-grams are proved to be strong features for stylistic analysis [7]. This representation is tolerant to grammatical errors, computationally cheap and applicable to various languages as it allows avoiding hard preprocessing (e.g. tokenization for oriental languages). A significant point in this approach is the choice of *N*. A larger *n* is able to capture contextual information and topic of the text, but it leads to dimensionality growth. A smaller *N* can capture subword information but fails to consider context. To reflect syntactical information, which is naturally useful for style determination, syntactic *N*-grams are presented in [8].

The writing style of a document is the essential evidence once causing the problem of author verification, responding the problem whether given documents have been written by the same author [9]. As usual, this task is limited here by the hypothesis that there is simply one applicant [10]. The problem of authorship verification can be considered as principal, because any author identification problem can be decomposed into a group of authorship verification problems [11].

3. Methodology

In this section, the proposed replica is presented. A new measure, named the Mean Dependency, is introduced to assess the association between a given text chunk and several earlier ones.

3.1 Mean Dependency

Let us consider $\mathbb{D}$ as a collection of finite-length documents such that each sub-document of any document belonging to $\mathbb{D}$ also belongs to $\mathbb{D}$, and take a semi-distance function *Dis* is defined on $\mathbb{D}\times\mathbb{D}$, where it is not suggested that *Dis*(D_1, D_2)=0 implies that D_1=D_2. In the framework of our model, we consider a document $\mathcal{D}\in\mathbb{D}$ a series of *m* sequential sub-documents: $\mathcal{D}=\{\mathcal{D}_1,...,\mathcal{D}_m\}$. In the formal language theory terminology, $\mathcal{D}$ is the concatenation of $\mathcal{D}_1,...,\mathcal{D}_m$. Our perception suggests that a document $\mathcal{D}$ is considered as an outcome provided by "a random number generator" reflecting the writing style of the authors. Aiming to quantify the evolution of a text within the writing process, we introduce the Mean Dependence characterizing the mean relationship between a chunk $\mathcal{D}_i$, *i=T+1,...,m* and the set of its *T* "precursors"

$$ZV_{T,Dis}(\mathcal{D}_i,\Delta_i)=\frac{1}{T}\sum_{\mathcal{D}\in\Delta_i}Dis(\mathcal{D}_i,\mathcal{D}),\tag{1}$$

where $\Delta_i=\left\{\mathcal{D}_{i-j},\, j=1,...,T\right\}$ the set of its *T* "precursors" of $\mathcal{D}_i$. To distinguish styles a new function measuring dissimilarity amongst texts pieces is proposed by the following way:

$$DZV_{T,Dis}(\mathcal{D}_i,\mathcal{D}_j)==\left|ZV_{T,Dis}(\mathcal{D}_i,\Delta_i)+ZV_{T,Dis}(\mathcal{D}_j,\Delta_j)--ZV_{T,Dis}(\mathcal{D}_i,\Delta_j)-ZV_{T,Dis}(\mathcal{D}_j,\Delta_i)\right|.$$

It is easy to see that it is also a semi-metric. Once $DZV_T(\mathcal{D}_i,\mathcal{D}_j)=0$ the sub-documents $\mathcal{D}_i$ and $\mathcal{D}_j$ exhibit close relationships with the own previous neighbors and the previous neighbors of another one. From the writing style standpoint the sub-documents appear to be very similar. The fact that DZV_T is actually not a metric might lead to ambiguous clustering process. To overcome this obstacle we provide a metrication of $(\mathbb{D},DZV_T)$ by means of the Fréchet-Kuratowski embedding (see, for example [12]) into the Euclidean space $\mathbb{R}^m$ equipped by the standard Euclidean norm $||\cdot||$ ($\pi:(\mathbb{D},DZV_T)\rightarrow(\mathbb{R}^m,||\cdot||)$:

$$\pi(x)=(DZV_T(\mathcal{D},\mathcal{D}_1),\ldots,DZV_T(\mathcal{D},\mathcal{D}_m)),\tag{2}$$

which induces a new metric on *D*:

$$DZVE_T(\mathcal{D}_1,\mathcal{D}_2)=||\,\pi(\mathcal{D}_1)-\pi(\mathcal{D}_2)\,||.\tag{3}$$

3.2 Distance Construction

Distance function choice is essential in the proposed approach. A relevant distance function may be extracted to reflect writing style attributes. It is a challenging and non-trivial task, which has been recently studied as a crucial subject. Formally, measures such as the Levenshtein distance (or the edit distance) can be applied. However, in the text mining domain it is more acceptable to convert texts into a probability distribution and afterwards to use a distance between them. In our context, we suggest that there is a transformation $\mathcal{F}$, which maps the documents belonging to $\mathbb{D}$ into the set $\mathbb{P}$ of the probability distributions on [0,1,2,…], and

$$Dis(\mathcal{D}_1,\mathcal{D}_2) = dis\left(\mathcal{F}\left(\mathcal{D}_1\right),\mathcal{F}\left(\mathcal{D}_2\right)\right),$$

where *dis* is a distance function (a simple probability distance) defined on $\mathbb{P}$. The probability metrics theory is stated in [13,14]. A comprehensive survey of distance/similarity measures between probability density is presented in [15]. In the current paper we use the following Spearman's correlation distance function:

$$Dis(\mathcal{D}_1,\mathcal{D}_2) = S(\mathcal{D}_1,\mathcal{D}_2) = 1-\rho(\mathcal{D}_1,\mathcal{D}_2),$$

where ρ is the Spearman's ρ (see e.g., [16]), which is calculated for distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ treated as a kind of ordinal data such that the frequency values are regarded as the rank positioning. This method has been successively applied to visual word histogram relationship evolution (see, for example [17]), and for clustering genomes within the compositional spectra approach [18]. As usual, a transformation $\mathcal{F}$ is constructed by means of the common Vector Space Model. This model disregards grammar and the order of terms but keeps the collection of terms. Each document is described via a terms frequency table in contradiction of the vocabulary containing all the words (or “terms”) in all documents in the corpus. The tables are considered as vectors in a linear space having a dimensionality equal to the vocabulary size. In the *N*-grams Model the vocabulary consists of all *N*-grams in the corpus. The *N*-gram approaches are widely applied in the text retrieval area (see, for example, [19]).

A Toy example

Let us suppose that the analysed corpus is based just on three *N*-grams, and two sets of four sequential documents given in the following tables with their occurrences and the ranks according to the occurrences. Table 3 presents the Spearman's ρ values calculated between first three documents of each set with the last ones.

	Occurrences				Ranks			
N-gram	$\mathcal{D}_1^{(1)}$	$\mathcal{D}_2^{(1)}$	$\mathcal{D}_3^{(1)}$	$\mathcal{D}_4^{(1)}$	$\mathcal{D}_1^{(1)}$	$\mathcal{D}_2^{(1)}$	$\mathcal{D}_3^{(1)}$	$\mathcal{D}_4^{(1)}$
1	5	8	10	20	3	2	1	1
2	10	3	0	8	2	3	3	2
3	15	10	3	2	1	1	2	3

Table 1. The first documents set.

	Occurrences				Ranks			
N-gram	$\mathcal{D}_1^{(2)}$	$\mathcal{D}_2^{(2)}$	$\mathcal{D}_3^{(2)}$	$\mathcal{D}_4^{(2)}$	$\mathcal{D}_1^{(2)}$	$\mathcal{D}_2^{(2)}$	$\mathcal{D}_3^{(2)}$	$\mathcal{D}_4^{(2)}$
1	7	2	10	10	2	3	2	2
2	3	3	20	25	3	2	1	1
3	9	4	5	4	1	1	3	3

Table 2. The second documents set.

N-gram	$\mathcal{D}_1^{(2)}$	$\mathcal{D}_2^{(2)}$	$\mathcal{D}_3^{(2)}$	$\mathcal{D}_1^{(1)}$	$\mathcal{D}_2^{(1)}$	$\mathcal{D}_3^{(1)}$
i=1	-1	-0.5	0.5	-0.5	0.5	1
i=2	-0.5	0.5	1	-1	-1	0.5

Table 3. Spearman's ρ with $\mathcal{D}_4^{(i)}$, $i = 1,2$.

So:

$$ZV_{3,\rho}\left(\mathcal{D}_4^{(1)},\left\{\mathcal{D}_1^{(1)},\mathcal{D}_2^{(1)},\mathcal{D}_3^{(1)}\right\}\right)=\frac{1}{3}(-1-0.5+0.5)=-\frac{1}{3},\ ZV_{3,\rho}\left(\mathcal{D}_4^{(2)},\left\{\mathcal{D}_1^{(2)},\mathcal{D}_2^{(2)},\mathcal{D}_3^{(2)}\right\}\right)=-0.5,$$

and

$$ZV_{3,\rho}\left(\mathcal{D}_4^{(1)},\left\{\mathcal{D}_1^{(2)},\mathcal{D}_2^{(2)},\mathcal{D}_3^{(2)}\right\}\right)=ZV_{3,\rho}\left(\mathcal{D}_4^{(2)},\left\{\mathcal{D}_1^{(1)},\mathcal{D}_2^{(1)},\mathcal{D}_3^{(1)}\right\}\right)=\frac{1}{3},\ DZV_{3,\rho}(\mathcal{D}_4^{(1)},\mathcal{D}_4^{(2)})=0.5.$$

4. Analysis of an Egyptian newspaper

The suggested methodology is appraised on editorial texts published in the Egyptian “Al-Ahraam” (“The Pyramids”) newspaper for the periods from 1.1.2010 to 31.1.2010, 1.1.2011 to 30.9.2011, and 1.1.2014 to 30.6.2014. These periods contain important events in the political life of the appropriate society, particularly, the “Arab Spring”, reflected by changes in the official ideology. The newspaper “Al-Ahraam” was founded in 1875 in Alexandria. It is the second oldest newspaper in Egypt and the most famed daily paper, not only in the country, but in the all Arab world. It comprises a wide range of problems ranging from politics and economy to sport and family issues and is, apparently, the most dominant newspaper in the Arab world. The stated results were reported earlier in [20]. In the first step, the stop-words are omitted, and the 3-grams having occurrences bigger than the ninth decile of the total 3-gram occurrences in a corpus were chosen. This quantity is 3961. Most of the remaining 3-grams arise just in a small portion with comparatively minor occurrences. After selecting the 3-grams demonstrating significant spread and variety of the 3-gram frequencies within the corpus, the following indicator is considered:

$$V_i=\frac{median(f_{ij},j=1,...,m)}{max(f_{ij},j=1,...,m)},$$

where f_{ij}=1,…,3961 is the (frequency of) occurrence of an 3-gram i in a document $j=1,...,m$. Note that here m=817. A 3-gram is accepted as frequent if its V_i value exceeds the 99th percentile. There are 34 such 3-grams. The following figure shows the graph of $ZV_{T,Dis}$ constructed for T=20.

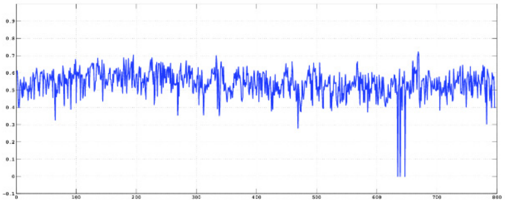

Figure 1: Graph of the *ZV-value*s constructed from the “Al-Ahraam” newspaper for *T*=20.

Two potential change points imply two changes in the newspaper style arising here. A clustering procedure is used in order to reveal the homogeneous periods in the social life reflected by the similar styles of the newspaper issues. The Partitioning Around Medoids (*PAM*) algorithm [22] is applied. In comparison to the *K*-means technique, *PAM* is a more robust procedure. *PAM* functions more appropriately for relatively small data due to its quadratic complexity regarding the number of items. As for most of the cluster partitioning algorithms, *PAM* includes the suggested number of clusters as its input parameters. “Smooth” clusters of the media documents rarely arise as a result of the in-built material noisiness. When trying to fix the problem, the attained partitions are smoothed using the wavelet approximation for level *L=6*, and the result is rounded to the nearest integer less than or equal to the value. In this approach the approximated signal is reconstructed merely using the approximation coefficients in the Haar wavelet decomposition. Clustering partitions are constructed for a growing from two numbers of clusters, and the process stops if the number of actual clusters decreases in a smoothed partition. This could suggest that the artificial clusters were omitted, and classifying with this or a bigger number of clusters is meaningless because it provides fake unsteady groups. In the studied problem the optimal number of clusters has been chosen as three. The corresponding partition with its smoothed version is presented in Fig 2.

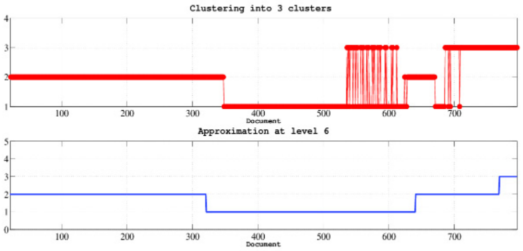

Figure 2: Partitioning into three clusters with the smooth version.

The smooth version presented in the bottom panel reveals three or four different styles. The first significant fluctuation point in the original (non-smooth) partition (one of 48 such points) appears at point 367 (2.1.2011). This position is suitably close to 25.1.2011, when many demonstrators congregated in Cairo's Tahir Square demanded the resignation of President Hosni Mubarak. A dense group of the oscillation points is located within positions 535-632, corresponding to the period from 9.6.2011 to 24.9.2011. The style here fluctuates between the "revolution" style defined by the "revolution" and an alternative new one. Apparently, such chaotic conduct is initiated by an unstable political situation. The next observable phase begins around position 690 (21.2.2014) and finishes at 714 (19.3.2014). Later, at point number 727 (1.4.2014), the style completely stabilizes. According to the Reuters publication, Egypt's government resigned on Monday (24.2.2014), paving the way for army chief Field Marshal Abdel Fattah al-Sisi to declare his candidacy for president of a strategic U.S. ally gripped by political strife. This date is really close to Friday 21.2.2014. On 24.3.2011, an Egyptian court condemned 529 followers of the Muslim Brotherhood to death (after 1.4.2014 the style completely stabilizes).

5. Authorship verification using strictly stationary model

In this section we discuss an application of the proposed methodology to the authorship verification problem. As was mentioned earlier, a document is treated as a consequence of "a random number generator" reproducing by the writing style of the document writers. It is very natural to presume that if a document is composed using the same personal writing style, then by an appropriate choice of the delay parameter *T*, the sequence $\{X_i = ZV_{T,Dis}(\mathcal{D}_i),$ $i = T+1,...m\}$ is a strictly stationary sequence of random variables. Subsequently, the inner inherent relationship between sequential parts of the document is kept. In other words, we assume that each X_i has the same probability distribution and, moreover, each finite set $\{X_{i_1+h}, X_{i_2+h}, ..., X_{i_n+h}\}$ has a joint distribution that does not depend on *h*. According to the Birkhoff ergodic theorem (see, for example, [22] Section 28.3) there exists a random variable *Y* such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = Y \tag{4}$$

is almost sure. When this sequence is ergodic, for instance having a summable covariance, *Y* is a constant. The discussed model makes it possible to offer the following authorship verification procedure. Let us suppose that we have two texts D_1 and D_2, which may have been written by the one author.

Algorithm: Verification if two documents are written by the same author (AV)
Input:

- $\mathcal{D}_1 \in \mathbb{D}$, $\mathcal{D}_2 \in \mathbb{D}$ - Two texts to compare.
- T - Value of the delay parameter T.
- TST - Two sample test procedure.
- L - Chunk size.

Procedure:

1. Divide document D_1 into m_1 chunks of size L : $D_1 = \{\mathcal{D}_1^{(1)}, ..., \mathcal{D}_{m_1}^{(1)}\}$.

2. Divide document D_2 into m_2 chunks of size L : $D_2 = \{\mathcal{D}_1^{(2)}, ..., \mathcal{D}_{m_2}^{(2)}\}$.
3. Concatenate the sequences:

$$D_0 = \{\mathcal{D}_1^{(1)}, ..., \mathcal{D}_{m_1}^{(1)}, \mathcal{D}_1^{(2)}, ..., \mathcal{D}_{m_2}^{(2)}\} = \left\{\mathcal{D}_1^{(0)}, ..., \mathcal{D}_{m_1+m_2}^{(0)}\right\}.$$

4. Calculate according to (1): $\mathbb{Z}_2 = \{ZV_{T,Dis}(\mathcal{D}_i^{(0)}), i = T+1, ..., m_1 + m_2\}$.
5. Calculate: $h = TST\left(\mathbb{Z}_1, \mathbb{Z}_2\right)$, where $\mathbb{Z}_2 = \{ZV_{T,Dis}(\mathcal{D}_i^{(0)}), i = T+1, ..., m_1\}$ and

$$\mathbb{Z}_2 = \{ZV_{T,Dis}(\mathcal{D}_i^{(0)}), i = m_1 + 1, ..., m_1 + m_2\}.$$

6. $if(h = 0)$ then $\mathcal{D}_1$ and $\mathcal{D}_2$, are written by the same author. Stop.
7. END
8. $if(h = 1)$ then $\mathcal{D}_1$ and $\mathcal{D}_2$ are not written by the same author.

Comments regarding the algorithm. Note that the key step of the procedure is a conversion of the considered texts into “time series” sequences performed in steps 1-3 of the algorithm. After this transformation the problem is actually converted into the change point detection problem. In our context, we are looking for a point separating the sequences generated by two documents by comparing the distributions of the sets.

$$\mathbb{Z}_{1,T,Dis} = \left\{ZV_{T,Dis}(\mathcal{D}_i^{(0)}), i = T+1, ..., m_1\right\}$$

and

$$\mathbb{Z}_{2,T,Dis} = \left\{ZV_{T,Dis}(\mathcal{D}_i^{(0)}), i = m_1 + 1, ..., m_1 + m_2\right\}.$$

Formally speaking, random variables within the series are not independent. So, a procedure in the spirit of the statistical inference for ergodic processes (see, for example, [23]) has to be formally applied. However, assuming in the framework of the model that the process is ergodic, due to (4) we conclude that

$$\lim_{T \to \infty} cov(Z_i, Z_j) = 0,$$

where $Z_i, Z_j \in \mathbb{Z}_{1,T,Dis}$ or $Z_i, Z_j \in \mathbb{Z}_{2,T,Dis}$. Thus, it appears to be reasonable to neglect the dependence between variables inside the series $\mathbb{Z}_{1,T,Dis}$ and $\mathbb{Z}_{2,T,Dis}$ for sufficiently large values of *T* and to compare distributions of $\mathbb{Z}_{1,T,Dis}$ and $\mathbb{Z}_{2,T,Dis}$ by means of a robust two-sample test.

Two-sample hypothesis testing is a statistical analysis approach designed to examine if two samples of independent random elements, drawn from the Euclidean space, have the same probability distribution function. For our purpose, such a procedure returns 1 if two samples have the same distribution and 0 otherwise. The most popular method is asymptotically distribution-free the Kolmogorov-Smirnov test (the *KS*-test). We use this test in our study.

The algorithm can be naturally generalized to provide a solution for the author recognition task. Let us suppose that there are several documents with known authorships $\left\{\mathcal{D}_1, ..., \mathcal{D}_n\right\} \in \mathbb{D}$, and we want to assign the document in question $\mathcal{D}_0$ to one of the authors. We can simply compare this document with all of $\left\{\mathcal{D}_1, ..., \mathcal{D}_n\right\} \in \mathbb{D}$ using the stated algorithm.

5.1 Numerical Experiments

In the texts involved in the experiments any uppercase characters are converted to the corresponding lowercase characters, and all other characters are unchanged. The stop-words are excluded, and only the 3-grams with occurrences over than 95th percentile are considered in the future. As mentioned earlier, we use the Kolmogorov-

Smirnov test with the traditional significance level threshold (0.05) for testing the statistical hypothesis. The results are presented in the tables of the statistical hypothesis testing: a value marked in bold indicates that the hypothesis of the same style is rejected. The experiments are provided through the chunk size *L=5000* with *T=20*.

5.1.1 Comparison of book collections

In the first stage several experiments are provided on large book collections.

- The first collection consists of the seven novels by Isaac Asimov's Foundation series (denoted as *F1,...,F7*): “Prelude to Foundation”, “Forward the Foundation”, “Foundation”, “Foundation and Empire”, “Second Foundation”, “Foundation's Edge” and “Foundation and Earth”.
- The second collection consists of the seven novels by Arthur C. Clarke (denoted as *AC1,...,AC7*): “2010: Odyssey two”, “2001: A Space Odyssey”, “A fall of Moondust”, “Against the fall of Night”, “Expedition To Earth”, “The Sands of Mars” and “The Wind From The Sun”.
- The third collection consists of the seven novels by Robert Sheckley (denoted as *RS1,...,RS7*): “The Mountain without a Name”, “Immortality, Inc”, “The Status Civilization”, “Mindswap”, “Options”, “The Laertian Gamble” and “Soma blues”.

The books in each one of the groups are compared to three files matching the collections and denoted *F0*, *AC0* and *RS0*. The 1091 most informative 3-grams are chosen according to the stated earlier criterion. The following figure demonstrates the graphs of $ZV_{T,Dis}$ values appearing here.

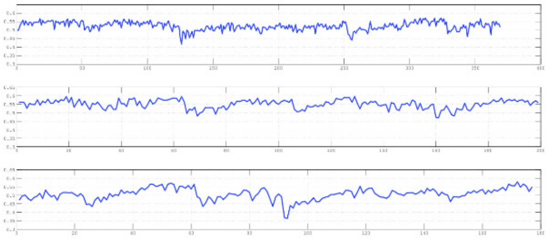

Figure 3: Graphs of *ZV-values* calculated for the *F0*, *AC0* and *RS0* collections.

Apparently, the charts present realizations of a stationary stochastic process. The averages and the standard deviations of the found values are 0.524, 0.550, 0.507; and 0.0244, 0.0257, 0.0347, correspondingly. It appears that the style of the second collection (*AC0*) is relatively more stable in comparison with two others due to its largest average value combined with small standard deviations. Table 4 exhibits the obtained *p-values*. Recall, the values indicating rejection of the null hypothesis are marked in bold. The obtained outcomes demonstrate absolute differences between the collections styles.

	F0	*AC0*	*RS0*
F0	0.9958	**4.934e-20**	**6.01e-12**
AC0	**4.934e-20**	0.9983	**3.066e-30**
RS0	**6.01e-12**	**3.066e-30**	0.9902

Table 4: $p-values$ in comparison of collections.

5.1.2 Authorship determination

In the next step several experiments are provided to evaluate the method's ability to discover authorship. To this aim we compare the following documents with the studied previously collections: “Rendezvous with Rama” by Arthur C. Clarke (denoted as *RR*), “Nemesis by Isaac Asimov” (denoted as *NEM*) and “A Call to Arms” by Robert Scheckley (denoted as *CA*). The proposed method success to recognize all writing styles:

	F0	*AC0*	*RS0*
RR	**8.045e-05**	0.390	**1.437e-06**
NEM	0.161	**1.570e-10**	**8.116-05**
CA	**0.0075**	**2.366e-08**	0.990

Table 5: $p-values$ in comparison of collections with three additional books.

5.2 Analysing of the Foundation Series

The Foundation series is the most popular science fiction series by Isaac Asimov with considerable impact in the nonfiction and the fiction fields. Three novels, “Foundation”, “Foundation and Empire”, “Second Foundation” are published in 1951, 1952 and 1953 correspondingly. Other four books are published in 1988, 1993, 1982, and 1986. It is very curious to investigate if such a considerable gap between the writing periods (about 30 years) can cause a significant change in the series style. To this aim the books series are compared. The *ZV-values* calculated for whole collection *F0* is given in Fig. 4.

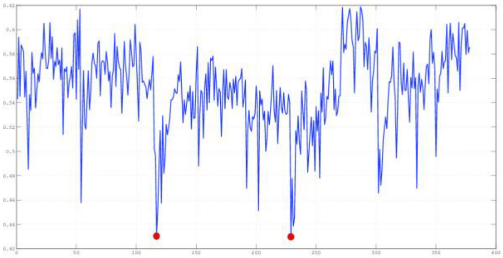

Figure 4: *ZV-values* calculated for the *F0* collection with possible “change points”.

Analogously with the discussion in Section 3 two possible “change points”, marked in red, may be seen. They are probably associated with the changes in the series style. Thus, a clustering approach can be used here to turn out novels similar from the style standpoint. The algorithm *PAM* is applied here in the manner stated earlier. We use the values of the *KS*-statistic obtained within the books comparison procedure and divide the *ZV-values* into clusters with the number of clusters running from 2 to 3. Note that there is no sense in considering a bigger number of clusters due to possible arising of small artificial stable partitions having, for example, groups composed just from one element.

To choose the most appropriate cluster solution amid those obtained, the silhouette criterion is applied. The silhouette [21] quantifies how much closer are points in a cluster in comparison to points belonging to other clusters. Items with large positive silhouette values around 1 are well clustered; those with negative silhouette values are positioned in a wrong cluster. The average of the silhouette values calculated for all points characterizes the partition quality. It is intuitively clear that a partition with the highest silhouette value is preferred because such a partition means it provides the best compact clusters separated as well as possible. However, the criterion is quite formal, and solutions obtained for numbers of clusters close to the optimal one may also be meaningful. To normalize the distances the diagonal element is subtracted from all elements in the corresponding row. According to the obtained values of the silhouette criteria, [0.4673,03104], two optimal number of clusters is chosen equal to two. The appropriate partition is presented in Tab. 3. The optimal clusters quantity is two.

book	cluster	year of publication
F0	**2**	**1988**
F1	**2**	**1993**
F2	1	1951
F3	1	1952
F4	1	1953
F5	1	1982
F6	**2**	**1986**

Table 6: Partitions of the *F0* collection into 2 clusters.

As can be seen, the obtained dichotomy of the Foundation Universe suits the writing periods of the books rather well. So, a significant gap between the writing periods (about 30 years) leads to this substantial difference in the styles.

6 Conclusion

The article suggests a novel methodology for modeling and classification of text documents. A new dynamic similarity measures between sequantial text chunks, named the Mean Dependency, is proposed in this paper. The evolution of this characteristic provides a time series representation of a document. Therefore, time series analysis methods can be inherently applied to text classification. Two possible implementations of the general methodology are considered in the paper. The first one deals with modeling and visualization of media. Change points appearing in the constructed time series can indicate fluctuations in the social life. The considered example of an Egyptian newspaper "Al-Ahraam" demonstrates the high ability of the proposed method to identify changes in the social life connected to the "Arab Spring" using only sequential issues of this daily paper. The second approach treats the appropriate time sequence as a strictly stationary sequence. The Kolmogorov-Smirnov two sample is applied here in order to check the significance of possible change points. In future research, we intend to involve other time series models to extend the methodology's application to as plagiarism detection in situation once plagiarism is expressed by changes in a text style. Another application may be dividing between real and artificially produced manuscripts.

References

1. Stamatatos, E. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 2009: 60(3): 538-556.
2. Juola, P. Authorship Attribution Foundations and Trends®, In *Information Retrieval*, 2008: 1(3): 233-334.
3. Burrows, J. "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing, 2002: 17(3): 267-287.
4. Jannidis, F., Pielstrom, S., Schoch, C. and Vitt, T., Improving Burrows Delta. In: Abstracts for the Digital Humanities 2015 Sidney. 10.
5. Oliveira, W., Justino, E., and Oliveira, L. S. Comparing compression models for authorship attribution. *Forensic science international*, 2013: 228(1): 100-104.
6. Cerra, D., Datcu, M., and Reinartz, P., Authorship analysis based on data compression. *Pattern Recognition Letters*, 2014: 42: 79-84.
7. Posadas-Duran, J. P., Sidorov, G., Batyrshin, I., Complete syntactic *N*-grams as style markers for authorship attribution. In *Human-Inspired Computing and Its Applications*, Springer International Publishing, 2014: 9-17.
8. Rhodes, D., Author Attribution with CNN, 2015 Reports - Stanford University.
9. Kestemont, M., Luyckx, K., Daelemans, W., and Crombez, T. Cross-genre authorship verification using unmasking. *English Studies*, 2012: 93(3): 340–356.
10. Luyckx, K., and Daelemans, W. Authorship attribution and verification with many authors and limited data. In *Proceedings of the twenty second international conference on computational linguistics (Coling 2008),* 2008: 513–520.
11. Koppel, M., and Winter, Y. Determining if two documents are written by the same author. *Journal of the American Society for Information Science and Technology*, 2014: 65(1): 178–187.
12. Morgan, C L. Embedding metric spaces in Euclidean space. *Journal of Geometry*, 1974: 5 (1): 101–107.
13. Zolotarev, V. Modern Theory of Summation of Random Variables. Modern Probability & Statistics Series. VSP; 1997.
14. Rachev, S. Probability metrics and the stability of stochastic models. Wiley series in probability and mathematical statistics: Applied Probability and Statistics. Wiley; 1991.
15. Cha, S H., Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences,* 2007: 1(4): 300–307.
16. Kendall, M G. and Gibbons, J D. Rank correlation methods. London: Edward Arnold; 1990.
17. Ionescu, R T., and Popescu, M. "PQ" Kernel. *Pattern Recognition Letters*, 2015; 55(C): 51–57.
18. Bolshoy, A, Volkovich, Z., Kirzhner, V., and Barzily, Z. Genome Clustering: From Linguistic Models to Classification of Genetic Texts ; vol. 286. Springer Science & Business Media; 2010.
19. Cavnar, W. B. and Trenkle, J. M. N-gram statistics for natural language understanding and text processing. In: *IEEE Trans. on Pattern Analysis* and Machine Intelligence, 1979: 2: 164–172.
20. Volkovich, Z, Granichin, O. Redkin, O. and Bernikova, O. Modeling and visualization of media in Arabic. *Journal of Informetrics*, to appear 2016.
21. Kaufman, L. and Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley; 1990.
22. Fristedt, B., and Gray, L. *A Modern Approach to Probability Theory*. Probability and Its Applications. Birkhäuser Boston; 1996.
23. Ryabko, D. and Ryabko, B. *Nonparametric statistical inference for ergodic processes . Information Theory, IEEE Transactions on* 2010: 56 (3): 1430–1435.