Relative N-Gram Signatures: Document Visualization at the Level of Character N-Grams

Magdalena Jankowska* Vlado Kešelj† Evangelos Milios‡

Faculty of Computer Science, Dalhousie University

ABSTRACT

The Common N-Gram (CNG) classifier is a text classification algorithm based on the comparison of frequencies of character n-grams (strings of characters of length n) that are the most common in the considered documents and classes of documents. We present a text analytic visualization system that employs the CNG approach for text classification and uses the differences in frequency values of common n-grams in order to visually compare documents at the sub-word level. The visualization method provides both an insight into n-gram characteristics of documents or classes of documents and a visual interpretation of the workings of the CNG classifier.

Keywords: Visual analytics, visual text analysis, text classification.

1 INTRODUCTION

A character n-gram from a given text is a string of n consecutive characters appearing in the text. Character n-gram based methods have been successfully applied to various problems in the text mining domain. They proved to be especially effective in the field of text classification. Language recognition [11] is the domain of one of the earliest, and very successful, applications of n-gram methods. Authorship attribution is another text classification problem that has been efficiently tackled by using character n-gram approach [9, 18, 33, 20, 24].

An important advantage of the character n-gram approach is its language independence. As the model is based on character sequences, it does not rely on syntax or semantics of a language; the same algorithm can be used for text documents of various languages. Stemming and removing of stop words is usually not necessary, and even not desirable, when a character n-gram method is used. Moreover, such an analytical method can be easily applied for languages of a non-trivial word segmentation. An n-gram does not even have to be a sequence of characters; an analogous methods as for text documents can be employed for other, non-character based sequences, for example in biological domain for classification based on amino acid sequences or DNA [32, 19, 34], or in the realm of music for authorship attribution task for musical works [36].

The foundation of our system is the Common N-Gram (CNG) analysis method: a text classification algorithm proposed by Kešelj et al. [24]. It relies on a dissimilarity measure between a document and a class that is based on the differences in the usage frequencies of the most common n-grams of the document and of the class.

We present a visualization system called Relative N-Gram Signatures that sheds light on n-grams used for the CNG analysis method. The system provides users with the ability to explore the most important (most distinguishing) n-grams for a document or for a class of documents as well as to gain insight into the inner workings of the classifier. The application employs visualization for discovery of patterns of characteristic n-grams, facilitating a visual inspection of n-grams that are characteristic for a given document or class. It also allows for a manual adaptation of the classification process via the visualization based on the task-dependant decision of a user. The system is targeted at users interested in detailed investigation of the characteristic n-grams of documents, for example the documents whose authorship is to be determined via CNG classification.

The rest of the paper is organized as follows. Section 2 provides background information on the CNG classifier and a survey of the literature related to our project. The main ideas we employ in our visualization are described in Section 3, which is followed by the presentation of the system capabilities and implementation in Section 4. We describe our experiments in Section 5 and conclude the paper by outlining plans for future work in Section 6.

*e-mail: jankowsk@cs.dal.ca

†e-mail: vlado@cs.dal.ca

‡e-mail: eem@cs.dal.ca

IEEE Conference on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

2 BACKGROUND

2.1 CNG classifier

The Common N-Gram (CNG) classifier is at heart of our Relative N-Gram Signatures. The classifier has been proposed by Kešelj et al. [24] and is based on comparing the frequencies of the most common character n-grams of the considered documents.

The character n-grams are strings of n consecutive characters from a given text. For example for the text "the table" there are 6 distinct 4-grams, namely "THE_", "HE_T", "E_TA", "_TAB", "TABL", "ABLE" (here the letters are converted to the upper case and the space is replaced with the underscore, as is the convention in this paper).

The CNG algorithm deals with the task of classifying a text document, that is with the task of labelling a document with a single label from a given fixed set of labels (or, in other words, of assigning the document to one class from a fixed set of classes). The authorship attribution is an example of a classification task: given a fixed set of author names the algorithm is to label a new document with one of these names. The classifier is built based on training data consisting of documents for which the class membership is know. For each class all training documents belonging to the class are concatenated together into one class document.

For each class and for a given document to be assigned to one of the classes, the classifier builds a *profile* that reflects the usage frequency of the most common n-grams of a given length. A profile is built based on the frequency of each n-gram in the corresponding document normalized by the length of the document (i.e., on the number of times the given n-gram appears in the document divided by the total number of n-grams of this length). Given the parameters L being the profile's size and n being the n-grams' length, a profile is the sequence of the L most common distinct n-grams of the length n, ordered by their decreasing normalized frequency.

For a pair of profiles P_1 and P_2 of two documents, the relative difference between frequencies of a given n-gram x, as proposed

in [24], is defined as follows:

$$d_x(P_1,P_2) = \left(\frac{f_{P_1}(x) - f_{P_2}(x)}{\frac{f_{P_1}(x)+f_{P_2}(x)}{2}} \right)^2, \quad (1)$$

where $f_{P_i}(x)$ is the normalized frequency of an n-gram x in the profile P_i, $i = 1,2$, with the assumption that $f_{P_i}(x) = 0$ whenever x does not appear in the profile P_i. The motivation is that rather than measuring an absolute difference in n-gram frequencies, which would give a too large weight for frequent n-grams, we measure relative difference. For example, a difference between frequencies $f_1 = 0.005$ and $f_2 = 0.003$ is the same as between $f_1 = 0.00005$ and $f_2 = 0.00003$, which is 0.25.

The total dissimilarity $D(P_1,P_2)$ between the two profiles is calculated by summing the distances between the profiles over all n-grams appearing in the union of the profiles:

$$D(P_1,P_2) = \sum_{x \in (P_1 \cup P_2)} d_x(P_1,P_2). \quad (2)$$

The CNG classifier applies the k-nearest neighbour algorithm with $k = 1$ based on the dissimilarity measure between profiles. For a new document to be classified, the dissimilarities between the profile of the document and the profiles of the classes (i.e., of the training class documents) are calculated and the new instance is assigned to the class with the least dissimilar profile.

2.2 Related work

A visual representation of classification algorithms aims at depicting the classification model, visually representing the classification results and facilitating better user interaction with the algorithm. Such visualizations have been proposed for some specific algorithms, such as Naïve Bayes [8], Decision Trees [7], or Support Vector Machines [14], or for classes of algorithms, such as additive classifiers [29], associative classifiers [12], or classifiers with probabilistic results [30].

The visualization of classification processes applied specifically to the text data poses its own specific challenges, especially due to the high dimensionality of the text representation, and has been less often reported. Nunzio [27] applies two-dimensional representation of a text document for probability-based classification and for visualization of text collections. Plaisant et al. [28] use Naïve Bayes for the text classification task as the basis of their interactive exploratory system for analyzing literary works by users that are not specialists in the text mining domain. The Ink Blots technique, proposed by Abbasi and Chen [5], is closely related to our approach in that it combines a classification model (namely a decision tree model based on an extensive set of text features) with a visualization of a document by superimposing the visual representation of these features over the analyzed text.

Our visualization system is based on another type of classification algorithm: the Common N-gram classifier relying on frequencies of character n-grams. The CNG method was originally applied to the authorship attribution [24]. The classifier has been also reported as a successful method for various other classification tasks: determination of software code's authors [18], genome classification [34], page genre classification [26], determination of composers of musical works [36], and recognition of computer viruses [6].

Other visualizations reported in the literature that are based on the n-gram similarity of documents (or sequences) aim at identification of clusters of related sequences.

Soboroff et al. [31] applied visualization for studying character n-grams as features for authorship attribution task. Their work is based on a different approach to this task: the authors apply Latent Semantic Indexing with character n-grams being the documents' terms. The authors visualize the documents by plotting their first LSI dimensions; it makes it possible to visually detect clusters of documents similar in terms of LSI components of n-gram terms.

Another type of a visualization employing character n-gram similarity of sequences is based on similarity matrices, where each column and each row corresponds to a sequence and the similarity between them is defined by means of the number of shared n-grams. This is the basis of a tool for the music visualization presented by Wolkowicz et al. [35]; the tool is based on a self-similarity matrix for a sequence encoding a musical piece and allows for detection of musical themes. Such an n-gram based similarity matrix is also used by Maetschke et al. [25] in a tool for visual comparison of sequences of biological domain (DNA, RNA, amino acid sequences), which enables users to identify clusters of related sequences.

Our system uses visualization for discovery of patterns of characteristic n-grams, which constitutes a way to visually depict and analyze text documents and their similarity.

Seasoft by Eick et al. [16] is one of the important examples of text visualization and analysis systems. It depicts various statistics related to lines of text to facilitate discovery of their patterns. Compus by Fekete and Dufournaud [17] is a system that allows for visual analysis of various XML atributes within a text. FeatureLens by Don et al. [15] integrates mining for frequent text patters with interactive visualisation for analysis of literary works.

A visual text analytic system that is probably most similar in flavour to our approach and that served as an inspiration for the way our viewer presents characteristic n-grams, is the Parallel Tag Clouds system by Collins et al. [13]. The authors visualize subsets of a faceted corpus by extracting words that distinguish a given subset and plotting these words as tag clouds in a parallel fashion. Our system aims at visualizing character n-grams instead of words. The feature of our system is its presentation of characteristic n-grams of a given background (base) document with respect to several other documents (and so it facilitates the discovery of patterns of n-grams that are distinguishing with respect to all analyzed documents or only with respect to some). It also allows for perception of the value of the distance in frequency for a given n-gram for a pair of documents.

Keim and Oelke [22] present a visualization technique called "literature fingerprinting" for investigation of how various literary analysis measures (such as the average sentence length or the vocabulary richness) vary between different parts of a text. The visualization allows for the perception of specific traits of authors that can be employed for the authorship attribution problem, and for gaining in-depth insight about characteristics of various parts of analyzed literature works. Our system is similar to this approach in the fact that it also facilitates the comparison of various text documents as a basis for authorship attribution analysis, but with an analysis based on character n-gram profiles. The different level of the visual analysis (the character n-gram level) and its relation to an underlying text classification method are the most important features differentiating our analysis method from the literature fingerprinting approach.

3 Visualizing characteristic n-grams

Our visualization of documents at the n-gram level is based on the idea of depicting the difference in the usage frequency of a given n-gram between two documents (where one of the documents may represent a class to which the other document may be assigned by a classifier). The goal of the visualization is to provide a user with an insight both into the characteristics of the documents in terms of chunks of words and into the inner workings of the CNG classifier.

3.1 Single relative signature

For a pair of documents we create a structure that we call a *relative signature* that reflects the difference of usage frequency of the most common n-grams between the documents. A relative signature of

two documents is built based on their n-gram profiles. Let P_1 and P_2 be the profiles of a given size L of two documents. Then, the relative signature is a sequence of all n-grams appearing in any of the profile, each of these n-grams coupled with the distance between the profiles with respect to the given n-gram. The n-grams are ordered as follows: first we include all the n-grams that appear in the first profile P_1 (the profile of the so-called base document, that serves as a "background" for the signature), ordered in the same way as in the first profile P_1, which are followed by the n-grams that appear only in the second profile P_2 (i.e., they do not appear in the first profile P_1), with preserving order as they appear in the second profile P_2. Thus, an n-gram with a number $k \leq L$ is the k-th most common n-gram in the base document, while n-grams with numbers greater than L appear only in the profile of the second document; in the latter case, the lower the number of an n-gram, the more common given n-gram is in the second document.

A visual relative n-gram signature is depicted in Figure 1. N-grams are represented by horizontal stripes. The n-grams are ordered from bottom up. For any n-gram the distance between the profiles with respect to the given n-gram is mapped to a colour according to a bipolar colour scale. The white colour indicates that the distance is close to zero i.e., the n-gram appears in both documents with a similar frequency. The red scale is used to encode a distance if the frequency of a given n-gram is higher in the base document (document with the profile P_1) than in the document with the profile P_2; the blue scale is used if the n-gram is less frequent in the base document than in the other document. The lighter a stripe is, the smaller the distance between profiles with respect to the corresponding n-gram.

One can notice that in the top part of the signature, presenting n-grams of the number higher than L, all stripes have the same darkest blue colour corresponding to the maximum distance; this is because these are the n-grams that do not appear in the base profile (and so the distance is the same for all of them and equal to its maximum value 4). Thus the blue top part of a signature conveys the information about how many n-grams appear in the profile of the second document only. For the same reason the bottom part of a signature, presenting n-grams of the number not higher than L, contains more red stripes than blue ones; that is because all the stripes of the darkest red colour, representing the n-grams that appear in the profile of the base document only, are located in this part of the signature.

As an example, in Figure 2 (that shows an enlarged part of the signature from Figure 1), it can be observed that among the first 19 most common 3-grams of *Alice's Adventures in the Wonderland* by L. Carroll, the n-grams "_SH" and "IT_" (where "_" denotes a sequence of non-letter characters) are more often used in *Alice's Adventures in the Wonderland* than in *Tarzan of the Apes* by E. R. Burroughs, while the n-gram "ED_" is used more often in *Tarzan of the Apes* than in *Alice's Adventures in the Wonderland*.

A total dissimilarity between two profiles, according to Equation 2, can be calculated by stepping over all n-grams in the relative signature of these two profiles and summing the corresponding distances.

A relative signature of a document on the background of itself (or, more generally, a relative signature of two documents with identical profiles) would be completely white (as the distance between profiles with respect to any n-gram would be zero) and would have the same size as the profiles (as there would be no n-gram that appears in only one of the profiles). In such a case, the dissimilarity between the (identical) profiles would be equal to zero. In general, the darker a signature is (i.e., the more of the dark stripes it includes, and the darker the stripes are), the higher the dissimilarity between the profiles is. Also the taller a signature is (i.e., the more of the n-grams appear in only one profile), the less similar the profiles are.

A single relative signature plot, when augmented with interaction features that allow a user to obtain information about a given

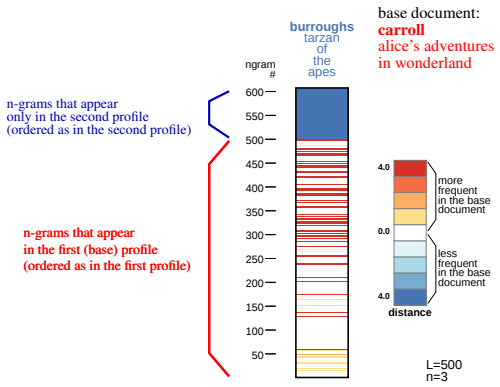

Figure 1: A relative signature of Burroughs' *Tarzan of the Apes* on the background of the base document of Carroll's *Alice's Adventures in Wonderland*, built on the profiles with the parameters $L = 500$ and $n = 3$.

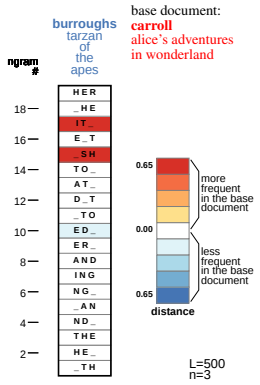

Figure 2: The first 19 n-grams (i.e., the bottommost 19 n-grams) of the relative signature of Burroughs' *Tarzan of the Apes* on the background of the base document Carroll's *Alice's Adventures in Wonderland*, built on the profiles with the parameters $L = 500$ and $n = 3$.

n-gram on demand, can be used for the exploration of the n-grams that are important for a given classification task.

3.2 Series of relative signatures

The possibility of determining patterns in the n-gram structures of several documents comes from plotting several relative signatures one next to another. For example, there may be a series of relative signatures of a given document with respect to various authors, or of relative signatures of a given author with respect to various documents. Such a visualization brings forth a potential of detecting interesting patterns, as n-grams that are characteristic for a document, when compared with various classes (authors) or n-grams distinguishing a class (an author) on a background of various documents. It also allows for a visual comparison of multiple entire relative signatures in order to compare the degree of similarity between documents.

The idea of a series of relative signatures is illustrated in Figure 3. Several relative signatures are plotted next to each other. Each of them has the same base document (Carroll's *Alice's Adventures in Wonderland*) that serves as a background for each signature. Thus L (here $L = 500$) bottom n-grams from each signature are the most common n-grams from *Alice's Adventures in Wonderland*. That means that for any number not higher than L, the n-gram of this number is identical in every signature. The n-grams with

numbers higher than L are specific to the second (not-base) document of a signature: these are n-grams that do not appear in the base profile, but appear in the profile of the second document. The relative signature of *Alice's Adventures in Wonderland* on the background of itself is plotted (as the second signature from the left) for reference.

Based on a series of signatures one can determine that among the five depicted books, Carroll's *Through the Looking Glass* has the most similar 3-gram usage statistics to *Alice's Adventures in Wonderland*. It is also evident that some characteristic 3-grams of *Alice's Adventures in Wonderland* differentiate this book from all others books depicted (a pattern of red or blue stripes across all signatures), while some 3-grams distinguish this book only from a subset of the depicted novels.

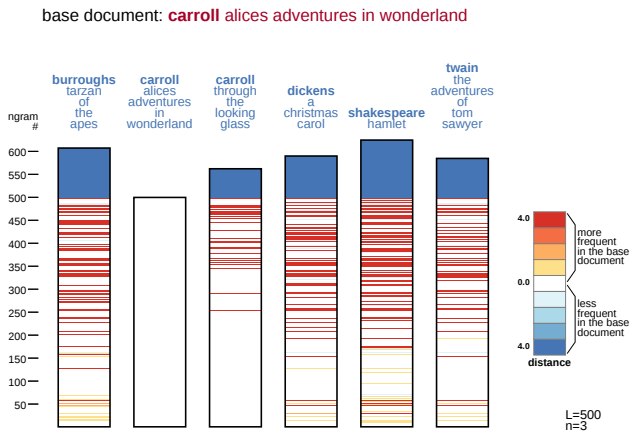

Figure 3: A series of relative signatures, each of them with the same base document: Carroll's *Alice's Adventures in Wonderland*, built on the profiles of 3-grams of the size 500.

4 Relative N-Gram Signatures

4.1 System capabilities

Our Relative N-Gram Signatures application is a visual text analytics system that enables users to draw and analyze relative n-gram signatures.

The sets of book profiles are prepared off-line. When n-grams are extracted, all letters are converted to uppercase and each sequence of non-letter characters (such as the space, punctuation marks or digits) is replaced by the underscore sign "_". Such treatment of characters—unifying n-grams that differ only in the usage of an uppercase letter as opposed to a lowercase one, or of a punctuation mark as opposed to the space—seems to be best suited for the visualization and the perception of patterns.

A user selects from a predefined set of documents the base document (that will serve as a background) and the documents to compare with the base document. The parameters of the profiles (the size L of the profiles and the length n of n-grams building each profile) are also defined by users. The selection is based on the predefined set of profiles prepared off-line; the currently available choices are $L = 500$ or $L = 800$, and n being of value 3, 4 or 5.

The set of relative signatures is prepared on-line based on the user's selection. The main view of the application is presented in Figure 4. A series of relative signatures is augmented by a bar graph illustrating the total dissimilarity between each document and the base document (bottom of the signature plot). The bar corresponding to the minimum dissimilarity has a distinguishing colour, which indicates to which class the base document has been assigned to by the CNG classifier.

A user can zoom into the signatures by double clicking on any signature, and zoom out by Shift double clicking. A button in the user interface allows for fast return to the default zoom level, i.e. to the level in which the entire signatures are visible. On a zoom-in level, two arrows provide the interface for browsing the signatures (moving up and down along the signatures). Figure 5 presents an example of the bottom part of the same set of relative signatures on two different zoom levels.

The area allocated for signature plots has a fixed height. The stripe width of each n-gram is based on this height and the number of n-grams in the longest signature, with a threshold minimum stripe width allowed. In a case when such an allocation would lead to the height of the longest signature plot exceeding the height of the plot area, the n-grams are sampled for plotting (the sampling is uniform with respect to the n-gram position in a signature, and identical in each signature). The sampling changes depending on the zoom level, i.e., after sufficient zooming each n-gram is plotted in any case.

When a mouse is held over a particular n-gram, a tooltip with the n-gram string and the distance between two given profiles with respect to this n-gram appears. Moreover, a highlighting horizontal bar is visible when a mouse hovers over an n-gram. This bar helps a user to analyze the given n-gram in all signatures. Naturally, the highlighting bar appears only when one of the first L n-grams is hovered. An example of a tooltip and the highlighting bar is visible in Figure 4.

Users may explore the context of n-grams in given documents on demand. On a mouse click on an n-gram the information about its context in both documents of the signature (the base document and the other document) is printed below the plot. The context information may be conveyed in a concordance style, that is by a list of examples of usage of the given n-gram with text fragments preceding and following the n-gram in the document (analogical to the "Key Word In Context" representation). Alternatively, the context information may be presented as a list of the most common words that contain the given n-gram, ordered according to their frequency in a document, and augmented by a bar plot illustrating this frequency. This representation provides a way to explore the link between n-grams and words; it allows for determining if there exists a meaningful pattern of words a given n-gram originates from, and what kind of pattern it is. Figure 6 presents two possible outputs a user can receive on a click on the n-gram "N_T_" in a relative signature on the background of *Alice's Adventures in Wonderland* by L. Carroll.

Some modification of the colour scale of the plot is also available for users. The maximum of the colour scale can be either set to the maximum of the distance over the entire signatures (which is the default option) or to the maximum of the distance over the parts of the signatures that are currently visible (after zooming in). Additionally, a user may choose to depict only the n-grams with the maximum distance (Figure 10 serves as an illustration of this option).

It is possible to search for n-grams in the visualization. A user specifies a text, which may be of the length n of the currently plotted n-grams, in which case it corresponds to a single n-gram, or may be longer. N-grams are extracted from the text and the user is provided with a list of those among them that appear in the visualized signatures. When the user selects a single n-gram from the list, the signature plots zoom into the area centred on the n-gram position and the corresponding n-gram stripes are highlighted.

Users may also modify the profile of the base document by removing n-grams that they deem to be not useful or even misleading for their analysis or classification task. This is performed by tagging on the plots the n-grams of the base document that are to be removed, and then re-running the analysis. In such a case a new profile of the base document is created: one that still has the re-

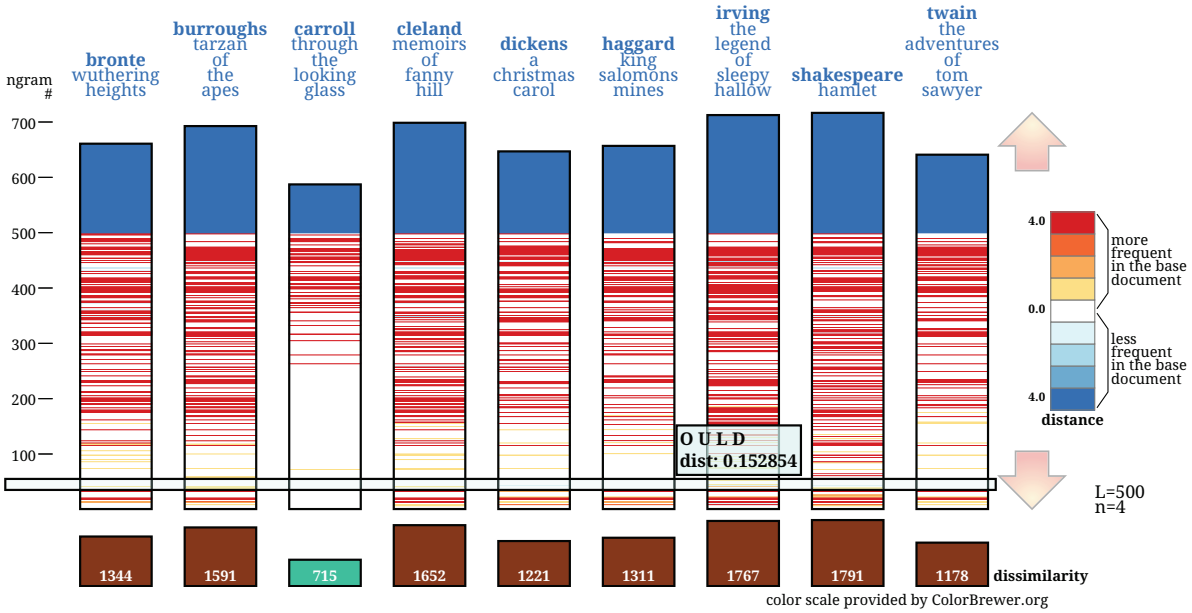

Figure 4: The main view of Relative N-Gram Signatures. The relative signatures of books by nine English authors with *Alice's Adventures in Wonderland* by L. Carroll as the base document, built on the profiles of 4-grams of the size 500.

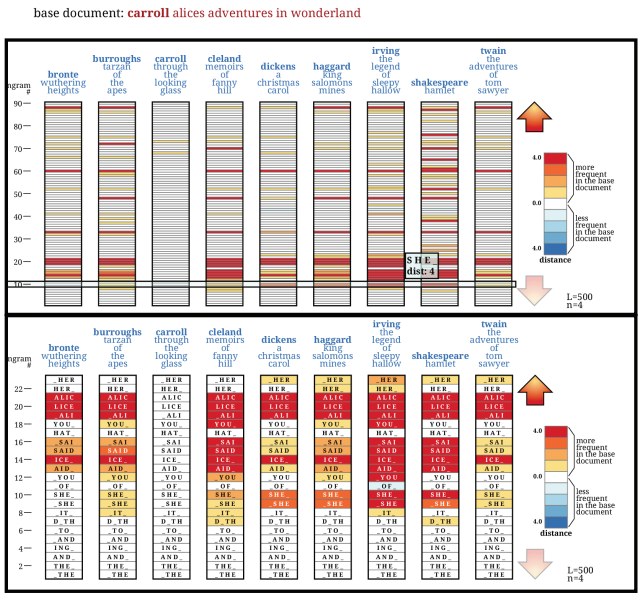

Figure 5: Two examples of zoom levels of the set of relative signatures depicted in whole in Figure 4.

quested length L but that does not contain the tagged n-grams; in other words k tagged n-grams are removed from the base document profile, and k new n-grams, with the frequency ranks $L+1,\ldots,L+k$, are added at the end of this profile. The profiles of the other documents are not modified. This process enables users to affect the visualization in a personal way, depending on their particular interests. It also affects the document dissimilarity values, and so may modify the classifier result.

Finally, the current plot can be downloaded by a user as a file in the SVG (Scalable Vector Graphics) format.

4.2 Implementation

The Relative N-Gram Signatures system is implemented as a web application.

The creation of profiles (the extraction of the most frequent n-grams and their normalized frequency values) is performed off-line. It is executed by an open-source Perl n-gram tool `Text::Ngrams` [23] by Vlado Kešelj (with an addition of a utf-8 support for non-ascii character n-grams).

The visualization module runs on the client side (the web browser side). It is developed using JavaScript visualization library d3.js by Bostock [10]. This library facilitates dynamic creation of SVG (Scalable Vector Graphic) elements that are bound with data.

Some of the user interactions trigger request to the server side. The actions that are performed on the server side are: the creation of relative signatures, the extraction of the context of an n-gram (performed by the regular expression search), the modification of the base document profile based on the user's n-gram removal, the extraction of n-grams to search for from the user provided text, and the creation of a plot file for downloading. These actions—executed by C++ and Perl programs—are called upon via AJAX (Asynchronous JavaScript and XML) request and PHP.

The colour scale has been provided by ColorBrewer.org [4].

5 Visual Analysis

In our experiments we used literature pieces in English and Polish. All books are in the public domain and have been downloaded either from the *Gutenberg* project [1] website or (for most of the polish texts) from the *Wolne Lektury* project [3] website. The only cleaning of these texts consisted of removing the parts at the beginning and/or the end of each book that relate to the above named literary projects.

5.1 Authorship attribution

The testbed for our system is based on the authorship attribution task. We use for our experiments the same set of books that were used by Kešelj et al. [24] for the original testing of the CNG classifier. This is a set of 12 English books by nine authors; for six

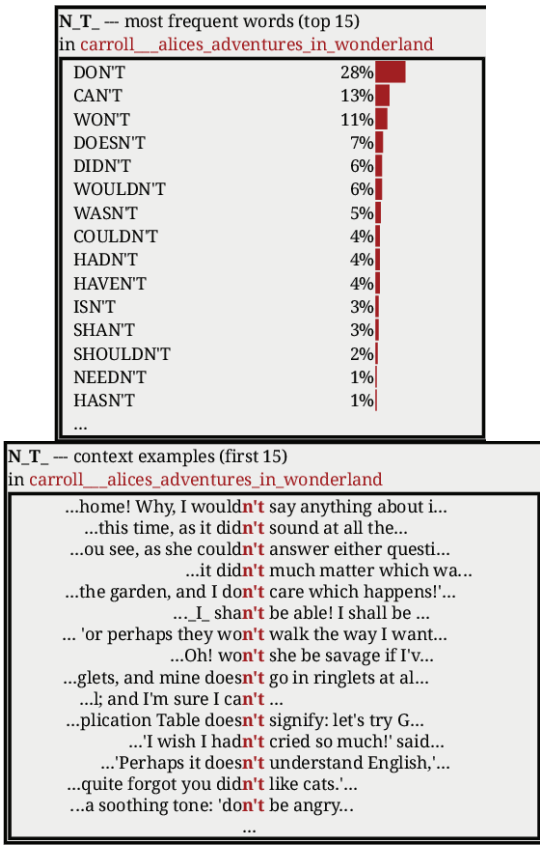

Figure 6: Two types of context information available for users, for the n-gram "N_T_" from the book *Alice's Adventures in Wonderland* by L. Carroll. The top panel presents the most frequent words that include this n-gram. The bottom panel shows the examples of the context of the n-gram in the concordance style.

authors there is one book in the set; for three authors there are two books in the set.

The relative signatures of nine books by nine authors on the background of the base document of *Alice's Adventures in Wonderland* by L. Carroll, are shown in Figure 4. One can easily distinguish the relative signature of *Through the Looking Glass* by L. Carroll that is much lighter and shorter than other signatures; this book represents the class to which *Alice's Adventures in Wonderland* has been (correctly) assigned by the CNG classifier.

Zooming into the signatures leads to discovering interesting patterns in terms of characteristic n-grams. Figure 7 shows a zoom-in level of the same signatures as the ones presented in Figure 4.

By hovering over interesting n-grams and clicking on them for the context information, one is able to gain interesting information. The n-grams #14, #19, #20 and #21 are the n-grams that originate most frequently from the word "alice" (these are n-grams "ICE_", "_ALI", "LICE" and "ALIC", respectively). They are used much more often in *Alice's Adventures in Wonderland* and in *Through the Looking Glass* (that is a sequel of *Alice's Adventures in Wonderland* with the same protagonist) than in any other of the analyzed books. Also characteristic for both books by Carroll are the n-grams "AID_", "SAID" and "_SAI" (originating mostly from the word "said"). These n-grams are very common in *Alice's Adventures in Wonderland*—as indicated by theirs position in the signatures (#13, #15, and #16, respectively)—and are used more often in the two books by Carroll than in the other books, with the maximum distance value for the books by Cleland, Irving, and Shakespeare. The n-gram "N_T_" (#48) has its source in the constructions such as "don't", "can't", "won't", etc. It serves as a distinguishing n-gram with respect to the books by Burroughs, Cleland, Haggard, Irving, and Shakespeare, but is used with a similar frequency by Bronte, Dickens and Twain (in the analyzed books). The n-gram #88 is "_VER" and is most often a chunk of the word "very"; it is used more commonly by Carroll than it is used in the other of the depicted books.

The blue stripes on the level of #128 correspond to the n-gram "_HE_". This n-gram comes from the word "he" and is used in the books by Carroll, Haggard and Shakespeare with a similar frequency, less frequently than in the other books.

It is also evident that all except for a couple of the 130 most frequent 4-grams of *Alice's Adventures in Wonderland* are used with a similar frequency in *Through the Looking Glass*. The first 4-gram of the *Alice's Adventures in Wonderland* profile that do not appear in the profile of its sequel is the n-gram "E_MO" (#134) that originates in *Alice's Adventures in Wonderland* most frequently from the words "the mock" (corresponding to the Mock Turtle, a character that appears only in *Alice's Adventures in Wonderland*).

It is interesting to see how n-grams related both to the writing style of an author ("N_T_", "_VER") and to the content of a book (for example, n-grams originating from the names of the characters in the novels) play their role in the decision of the CNG classifier.

Obviously, not every case provides such a clear distinction between relative signatures as in Figure 4. As an example, Figure 8 presents relative signatures of nine books by nine authors on the background of the base document of *A Christmas Carol* by C. Dickens. The base document is correctly assigned by the classifier to the other book by Dickens. The relative signature of *A Tale of Two Cities* by Dickens demonstrates the higher similarity of this book to the base document than the similarity of the other books, but the difference between this signature and the relative signature of *The Adventures of Tom Sawyer* is not easily visually perceived.

One can notice that the visualization of relative n-gram signatures does not only serve the purpose of facilitating discovery of characteristics of a document, but it also provides visual representation of the "reasons" of a classifier decision (a visual answer to the question "Why this book has been assigned to this author?" that is easier to interpret than an answer that consists of a single numeric dissimilarity value).

5.2 Mark Twain's novels

Our other experiment uses a set of novels by Mark Twain. The inspiration for this experiment comes from the research of Keim and Oelke [22] on the visual investigation of literary analysis measures. These authors' visualization demonstrates, among others, how much one of the novels of Mark Twain, *Adventures of Huckleberry Finn*, stands out from the other works of the writer with respect to such literary analysis measures like function words frequency, Simpson's index, and Hapax Legomena. We pointed our Relative N-Gram Signatures at the same set[1] of nine novels by Mark Twain to examine the difference between *Adventures of Huckleberry Finn* and other books by Mark Twain at the level of character n-grams.

Figure 9 presents the relative signatures of nine novels by Mark Twain on the background of the base document that is a concatenation of all of these books, with the parameter n set to 5 and L set to 500. One can observe that the signature of *Adventures of Huckleberry Finn* with respect to this *all-in-one* text stands out from the other ones (its profile is most dissimilar).

In particular, there is a lot of n-grams very common in general in Mark Twain's writing (depicted in the bottom of the signatures)

[1] We decided not to include one of the books by Mark Twain from this set though: we excluded *The Gilded Age* because it is co-authored with another writer.

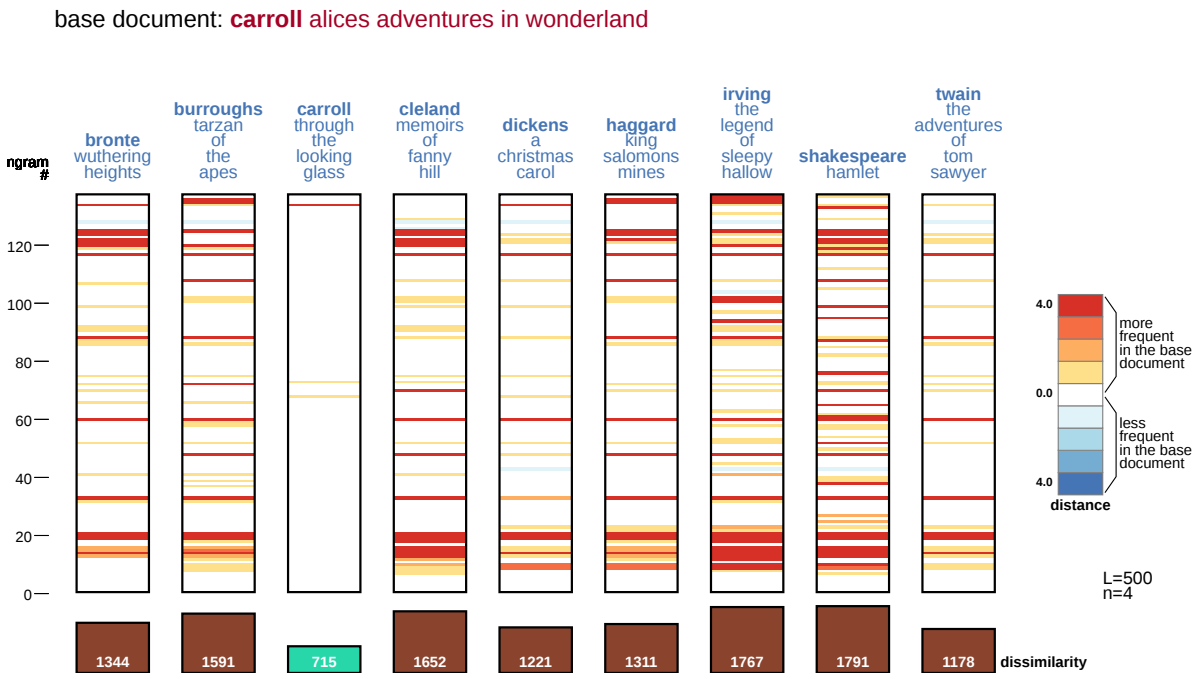

Figure 7: Relative signatures of nine books by nine English authors on the background of the base document of *Alice's Adventures in Wonderland* by L. Carroll, built on the profiles of 4-grams of the size of 500. The figure presents a zoom-in level, with around 130 first n-grams visible.

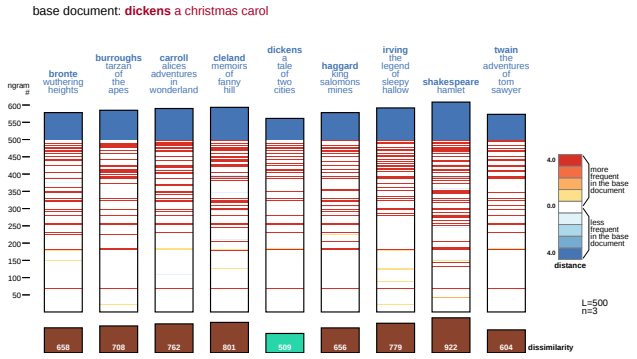

Figure 8: Relative signatures of nine books by nine English authors on the background of the base document of *A Christmas Carol* by C. Dickens, built on the profiles of 3-grams of the size of 500.

that are used with quite similar frequency in all of these nine books with the exception of *Adventures of Huckleberry Finn*. By closer examination of the signatures one can determine what these interesting n-grams are, and from which words they originate. These n-grams include, among others, "TION_" (#42) that is used much less often in *Adventures of Huckleberry Finn* than in all other books and that comes mostly from the words "attention", "question", "nation", "condition", "population", etc. This can possibly be interpreted as the less frequent usage of "formal" vocabulary in the book being a narrative of a 13-year old Huckleberry Finn. Other interesting examples are the n-grams "_WERE" and "WERE_" (#59, #60), that come mostly from the word "were" or the n-grams "_BEEN" and "BEEN_" (#181, #182) that have their source most frequently in the word "been"; these n-grams are used with a similar frequency in all the books with the exception of *Adventures of Huckleberry Finn* where they appear much less often. That suggests different grammar constructions used in the latter book. Other stylistic choices characteristic for *Adventures of Huckleberry Finn* seem to be indicated for example by the less frequent appearance in this particular book of the n-gram "D_NOT" (#175), originating in Mark Twain's texts mostly from the phrases "did not", "could not", "would not", or by the more frequent usage of the n-grams "_DOWN" and "DOWN_" (#172, #179) having its source mostly in the word "down", and of the n-gram "DN_T_" (#277), coming from the constructions "didn't", "couldn't", "wouldn't", "hadn't".

It is also apparent that the book *The Adventures of Tom Sawyer* shares some of the characteristics of *Adventures of Huckleberry Finn*. For example only in these two books the n-grams "WHICH", "_WHIC", and "HICH_" (#116, #117, #119) that have their source most frequently in the word "which", as well as the n-gram "T_IS_" (#111) that originates usually from the words "it is", "that is" and "what is", and the n-gram "WILL_" (#216) corresponding to the word "will" are used less frequently than in the base document. On the other hand, the n-gram "_IT_S" (#332), originating mostly from the phrase "it's", is more frequent in these two books than in the other Mark Twain's novels.

While *Adventures of Huckleberry Finn* have evidently the most distinguishing relative signature, *The Adventures of Tom Sawyer* and *The Prince and the Pauper* have many more characteristic n-grams than the other six books, which have rather uniform relative signatures.

5.3 Authorship attribution for Polish novels

An important feature of the character n-gram approach to the text analysis is its language independence. Relative N-Gram Signatures can be pointed at documents in any language provided the text is encoded in the utf-8 (Unicode Transformation Format-8) standard. The illustrative example is our analysis of a set of books by Polish authors. Figure 10 presents relative signatures of eight Polish books by eight authors, with the base document being another book by one

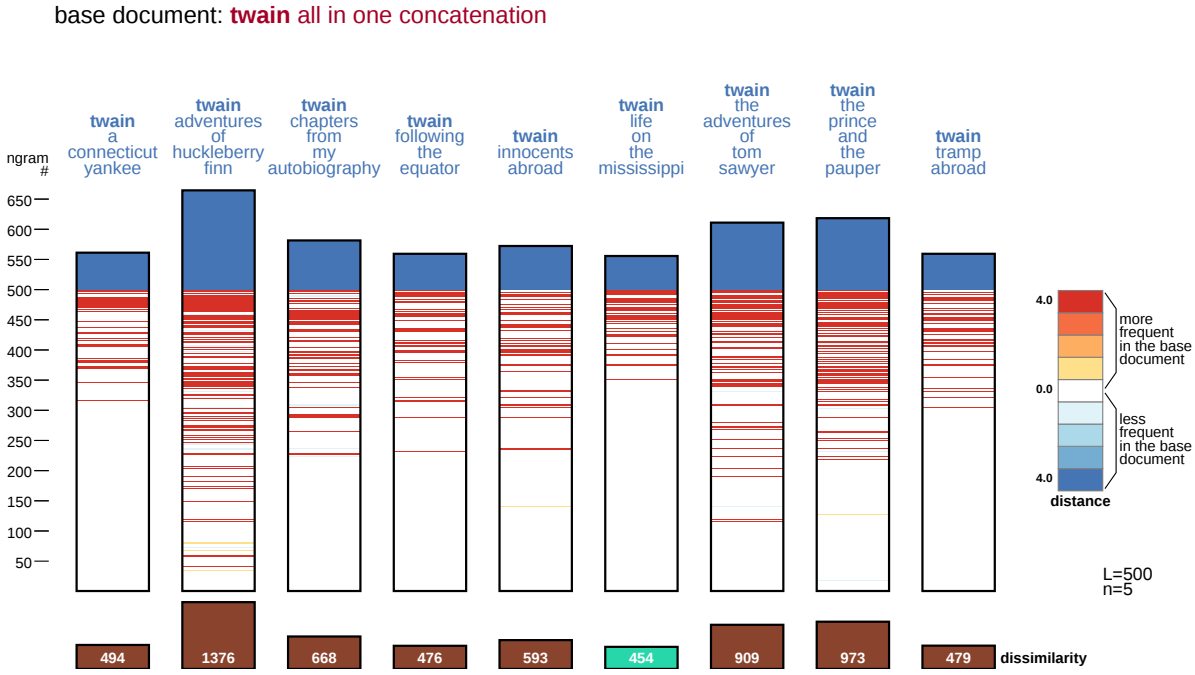

Figure 9: Relative signatures of nine books by Mark Twain with the concatenation of all these books as the base document, built on the profiles of 5-grams of size of 500.

of them—a novel by Henryk Sienkiewicz (correctly assigned to this author by the classifier). It is interesting to investigate the characteristic n-grams in depth and observe how the CNG classifiers captures characteristic words regardless of the many grammatical forms they assume due to the highly inflective nature of Polish. For example, one can discover that the n-gram “ZEKŁ” (#39) is characteristic for these two analyzed novels by Henryk Sienkiewicz. By investigating the words this n-gram is originating from, one discovers many forms of the verb “rzec” (a stylistically marked verb meaning “to say”, being characteristic for this writer) such as “rzekł”, “rzekła”, “rzekłszy” , “rzekłbyś”, “rzekłem” that are various grammatical forms corresponding to different grammatical persons, moods or to a participle form of this verb. Similarly, several different inflected grammatical forms of the noun "książę" (meaning “prince”) such as “książę”, “księcia”, “księciu” are the source of the n-gram “_KSI” (#45), characteristic for some of the analyzed books.

5.4 Subjective modification of the visualization

Users may modify the profile of the base document by subjective decision which n-grams not to include in the analysis.

We analyze an example of a difficult authorship attribution task from the Ad-hoc Authorship Attribution Competition, 2004 [2], [21]. Problem G of the contest presents the participants with four testing documents, each to be attributed to one of two authors (Author 01 or Author 02). For each of these authors there are three documents labelled by the author given for the purpose of training the classifier.

We point our system at the classification of the testing document sample 02 from this problem, with the CNG parameters n=4 and L=500. The other documents to compare this base document with are the documents obtained by the concatenation of the given training documents by Author 01 and Author 02, respectively.

The classifier attributes the sample to Author 02, as shown in Figure 11(a). By zooming in we can clearly see a set of very common n-grams used in the sample, that are used less often in the given works of these authors, and that originate from the word "Tarzan"

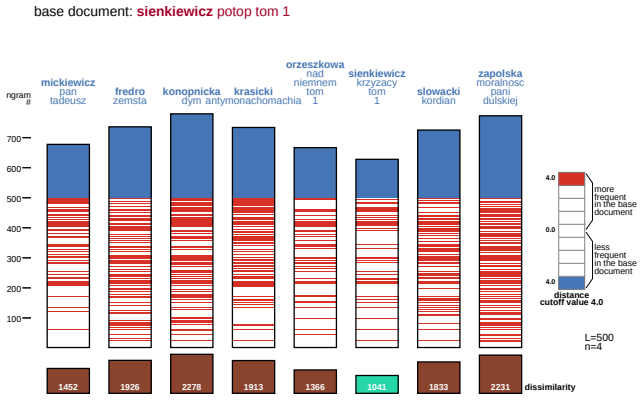

Figure 10: Relative signatures of eight Polish books by eight authors on the background of another book by one of these authors—Henryk Sienkiewicz, built on the profiles of 4-grams of size 500. Only the n-grams with the maximum distances are depicted.

(as one can discover by exploring the context of these n-grams in the analyzed documents). One can observe that the word "Tarzan" appears in fact in all three analyzed documents, with different frequencies. By browsing the signatures further, and by exploring the context of characteristic n-grams, one is able to find other n-grams that originate from proper names, namely from the words "Clayton", "D'Arnot", and "Jane". These n-grams are visually discovered because they are used with a higher frequency in the base document. The search capability allows a user to find all 4-grams that originate from these words, and the user can check that each of them except for the 4-gram "NOT_" comes mainly from these proper names (the 4-gram "NOT_" originates primarily from the word "not" and only secondarly from "arnot").

This may lead the user to the decision of removing the

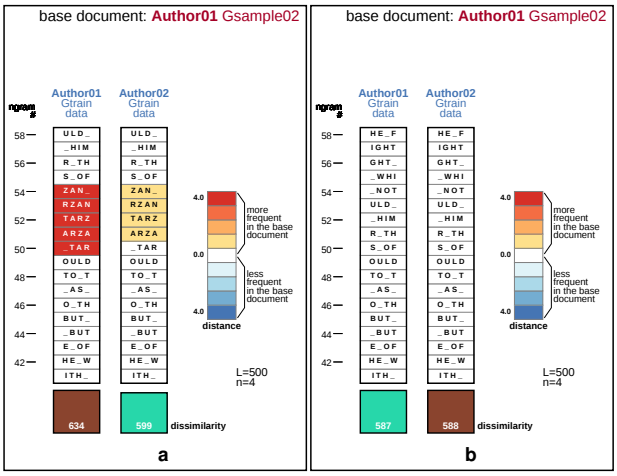

Figure 11: An example of influencing a classifier by modifying the profile of the base document. A zoom-in level of the signatures of two documents by two authors on the background of a sample document to be classified (Ad-hoc Authorship Attribution Competition, 2004, Problem G), before (a) and after (b) a manual removal of 19 n-grams from the profile of the base document.

following 19 n-grams from the profile of the base document: "_TAR", "TARZ", "ARZA", "RZAN", "ZAN_", "_CLA", "CLAY", "LAYT", "AYTO", "YTON", ,"TON_", "_D_A", "D_AR", "_ARN", "ARNO", "RNOT", "_JAN", "JANE", "ANE_", as their presence and the differences in the usage frequencies are judged to be the evidence of persons that the given documents are about, not the evidence of the difference in the style of writing, which is the topic of the analysis. The removal of these n-grams from the profile of the base document means that 19 new n-grams are added to this profile (which can be thought of as, though is not equivalent to, creating a profile of 500 most common n-grams of the base document stripped of these discovered proper names). After re-running the analysis the bottom part of the signatures is cleared of some red stripes (red indicating that the n-grams are used more often in the base document) while the classification result changes: the sample document is now attributed to Author 01 (with a small difference in the dissimilarity value: 587 vs 588), as shown in Figure 11(b).

This modified result is in fact the correct one: the test document 02 is a sample of Author 01 (the writings of Author 01 and Author 02 being respectively the early (pre-1914) and late (post-1920) writings of Edgar Rice Burrows [21]).

The method of modifying the profile of the base document based on a particular, task-dependent interest of a user is of a general use for visual analysis of the characteristics of documents (it can serve the purpose of removing from the visualization the elements deemed to be not useful or misleading in order to concentrate one's attention on other characteristics). As a method to modify the result of the classification algorithm it may be limited to such specific (and difficult) cases; more experiments are needed to evaluate that. This process can increase the trust of expert users in the decision of the classifier, because it allows them to eliminate at least some input to the decision from features that are in the users' opinion not relevant to the task.

6 Future Work

There is a number of possible additions to the current framework that can be made. The function of the distance between profiles with respect to a given n-gram that is currently used (Equation 1) is one of many that can be tested in this context (some already proposed in the literature [34], [18]). It would be of interest to compare different distance functions (with the selection being available for the system users) with respect to both the correctness of the classifier (for example for the authorship attribution problem) and the effect on the visualization.

We also plan to investigate ways of improving the capability of analyzing long signatures. In the case of other distance functions it may be possible to add an "overview mode" of a visual relative signature, namely a plot of a signature with only some fixed number of the most distinguishing n-grams (the n-grams with the highest distance) depicted, which would help to analyze longer profiles. For the currently used distance formula this mode does not limit the number of depicted n-grams significantly, as all of the quite numerous n-grams that appear only in one profile bear the same, maximum value of the distance. An alternative view in which n-grams are ordered by their distance value in each signature independently is also worth exploring as a way of comparing long signatures.

Further research on the ways of selecting the n-grams for profiles is another direction worth exploring. Comparing frequencies of n-grams across documents during the stage of building a profile (by use of some statistics as χ^2 or G^2) is one possibility; selection of variable length n-grams is another. Yet another approach worth investigating is to differentiate between n-grams that come from different parts of speech.

We are planning to experiment with pointing our system at different data sets and classification tasks, such as for example the task and the data related to the classification of web pages according to their genre. Usage of this tool for a problem of clustering text documents is another extension worth pursuing.

7 Conclusion

We presented Relative N-Gram Signatures, a visual text analytics system based on the Common N-Gram classifier. The system enables users to gain insight into the patterns of characteristic n-grams of given documents (with a relation to a set of documents) as well as to visually analyze the classifier algorithm. It also facilitates influencing the classification process by a user based on the insight gained from the visualization. We have demonstrated how this analysis can be performed by using a set of English literary novels and with the analytical task of analyzing authorship style. The robustness and language independence of the method is demonstrated by using a set of Polish novels in a similar analysis. No specific language knowledge is used. A particular advantage of using character n-grams has been shown to be robustness in the context of a highly inflective language, such as Polish.

The presented visualization tool is made available through any web browser by using JavaScript library d3.js, and it is very platform independent and relying only on availability of a recent version of a Web browser. The interactive feature of zooming and visualizing n-gram context provides a support for drill-in based analysis of documents.

Acknowledgements

The authors thank The Boeing Company for their support.

References

[1] Gutenberg project. http://www.gutenberg.org/, accessed on Jan 10, 2012.

[2] Ad-hoc authorship attribution competition, 2004. http://www.mathcs.duq.edu/~juola/authorship_contest.html, accessed on June 25, 2012.

[3] Wolne lektury project. http://www.wolnelektury.pl/, accessed on Mar 15, 2012.

[4] Colorbrewer 2.0. http://colorbrewer.org/, accessed on Nov 1, 2012.

[5] A. Abbasi and H. Chen. Categorization and analysis of text in computer mediated communication archives using visualization. In *Pro-*

ceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, JCDL'07.

[6] T. Abou-Assaleh, N. Cercone, V. Kešelj, and R. Sweidan. Detection of new malicious code using n-gram signatures. In *Proceedings of the second Annual Conference on Privacy, Security, and Trust, PST'04*, Fredericton, New Brunswick, Canada, October 2004.

[7] M. Ankerst, M. Ester, and H.-P. Kriegel. Towards an effective cooperation of the user and the computer for classification. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'00*.

[8] B. Becker, R. Kohavi, and D. Sommerfield. Visualizing the simple bayesian classifier. *Information visualization in data mining and knowledge discovery*, 18:237–249, 1997.

[9] W. R. Bennett. *Scientific and Engineering Problem-solving with the Computer (Prentice Hall series in automatic computation)*. Prentice Hall, first edition edition, 1976.

[10] M. Bostock. d3.js javascript library. `http://mbostock.github.com/d3/`, accessed on Nov 20, 2011.

[11] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, SDAIR'94*, pages 161–175, 1994.

[12] D. Chodos and O. R. Zaiane. Arc-ui: Visualization tool for associative classifiers. In *Proceedings of the 12th International Conference Information Visualisation, IV'08*.

[13] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze facted text corpora. In *Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology, VAST'09*, pages 91 – 98, October 2009.

[14] D. Cook, D. Caragea, and V. Honavar. Visualization for classification problems, with examples using support vector machines. In *Proceedings of the COMPSTAT 2004, 16th Symposium of IASC*, 2004.

[15] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the 6th ACM conference on Conference on Information and Knowledge Management, CIKM'07*.

[16] S. G. Eick, J. L. Steffen, and E. E. S. Jr. Seesoft - a tool for visualizing line oriented software statistics. *IEEE Transactions on Software Engineering*, 18:957–968, 1992.

[17] J.-D. Fekete and N. Dufournaud. Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *Proceedings of the 5th ACM conference on Digital Libraries, DL'00*.

[18] G. Frantzeskou, E. Stamatatos, S. Gritzalis, and S. Katsikas. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th international conference on Software engineering, ICSE'06*, pages 893–896, 2006.

[19] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and J. Klein-Seetharaman. Comparative n-gram analysis of whole-genome protein sequences. In *Proceedings of the second international conference on Human Language Technology Research, HLT'02*, pages 76–81, 2002.

[20] J. Houvardas and E. Stamatatos. N-gram feature selection for authorship identification. In *Proceeding of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, AIMSA'06*, pages 77–86, September 2006.

[21] P. Juola. Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334, Dec. 2006.

[22] D. A. Keim and D. Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology, VAST'07*, pages 115–122, 2007.

[23] V. Kešelj. Perl package text::ngrams. `http://www.cs.dal.ca/~vlado/srcperl/Ngrams`, accessed on Feb 1, 2012.

[24] V. Kešelj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, pages 255–264, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003.

[25] S. R. Maetschke, K. S. Kassahn, J. A. Dunn, S.-P. Han, E. Z. Curley, K. J. Stacey, and M. A. Ragan. A visual framework for sequence analysis using n-grams and spectral rearrangement. *Bioinformatics*, 26:737–744, March 2010.

[26] J. E. Mason, M. Shepherd, J. Duffy, V. Kešelj, and C. Watters. An n-gram based approach to multi-labeled web page genre classification. In *Proceedings of the 43rd Hawaii International Conference on System Sciences, HICSS'10*, pages 1–10, Hawaii, January 2010.

[27] G. M. D. Nunzio. Using scatterplots to understand and improve probabilistic models for text categorization and retrieval. *International Journal of Approximate Reasoning*, 50(7):945 – 956, 2009. <ce:title>Special Section on Graphical Models and Information Retrieval</ce:title>.

[28] C. Plaisant, J. Rose, B. Yu, L. Auvil, M. G. Kirschenbaum, M. N. Smith, T. Clement, and G. Lord. Exploring erotics in emily dickinson's correspondence with text mining and visual interfaces. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, JCDL'06*.

[29] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, B. Pearcy, C. MacDonell, and J. Anvik. Visual explanation of evidence in additive classifiers. In *Proceedings of the 18th conference on Innovative Applications of Artificial Intelligence, IAAI'06 - Volume 2*.

[30] C. Seifert and E. Lex. A novel visualization approach for data-mining-related classification. In *Proceedings of the 13th International Conference Information Visualisation, IV'09*, pages 490–495, 2009.

[31] I. M. Soboroff, C. K. Nicholas, J. M. Kukla, and D. S. Ebert. Visualizing document authorship using n-grams and latent semantic indexing. In *Proceedings of the 1997 workshop on New paradigms in information visualization and manipulation, NPIV'97*, pages 43–48, 1997.

[32] V. V. Solovyev and K. S. Makarova. A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. *Computer applications in the biosciences : CABIOS*, 9(1):17–24, February 1993.

[33] E. Stamatatos. Author identification using imbalanced and limited training texts. In *Proceeding of the 18th International Workshop on Database and Expert Systems Applications, DEXA'07*, pages 237–241, September 2007.

[34] A. Tomovic, P. Janicic, and V. Kešelj. n-gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer Methods and Programs in Biomedicine*, 81:137–153, February 2006.

[35] J. Wolkowicz, S. Brooks, and V. Kešelj. Midivis: Visualizing music pieces structure via similarity matrices. In *Proceedings of the 2009 International Computer Music Conference, ICMC'09*, pages 53–6, Montreal, Quebec, Canada, August 2009.

[36] J. Wolkowicz, Z. Kulka, and V. Kešelj. n-gram based approach to composer recognition. *Archives of Acoustics*, 33(1):43–55, January 2008.