# A Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis

## Notebook for PAN at CLEF 2014

Satyam, Anand, Arnav Kumar Dawn, and Sujan Kumar Saha*

Department of Computer Science and Engineering
Birla Institute of Technology, Mesra, Ranchi, India - 835215
{satyam.sinha.ds, anand62002, arnavdawn58, sujan.kr.saha}@gmail.com

**Abstract** In this paper we present a Latent Semantic Analysis (LSA) based approach to the Authorship Identification task in the PAN workshop. We apply LSA on a character n-gram based statistical model to obtain the similarities between document pairs. A statistical analysis of the pairwise document similarities is then used to determine a threshold value. Finding the optimal parameters (n-gram lengths, SVD cutoff, local and global weighting schemes), distance measures and thresholds for different languages and genre is an exercise central to this technique. The vastness of the parameter space makes the technique very flexible. The modularity of the technique allows easy tweaks to a specific module without requiring changes in other modules. This approach has very low runtime as it does away with the typical pre-processing and calculations involved in linguistic analysis.

## 1 Introduction

Authorship of documents is often questioned in legal systems or in the case of newly discovered manuscripts or articles of great literary or scientific value. The task of Author Identification [6], as presented in the PAN-2014 workshop, requires us to verify if a disputed document is the work of a given author, provided that other sample documents written by the author are available for analysis. Automation of this task has been of much interest to the research community as it is sought to mimic the wisdom of the linguistic experts by creating systems that can recognize unique elements (grammatical structure, stylistic markers) in the documents and attribute or rescind the authorship. The task is inherently tedious, requires a vast knowledge and experience in stylometry and moreover the human expertise in this area is in short supply. Hence, there is a need for computationally efficient systems that can verify authorship with reasonable accuracy.

It is desirable for such systems to work across different languages and genres. In the PAN-2014 workshop, the corpus used to evaluate the performance of the systems included texts written in languages like Dutch, English, Greek and Spanish with genre varying between articles, essays, novels and reviews.

* Faculty Advisor to the project

## 2 Motivation and Objectives

We aspire to create a general method that can adapt to different languages or genre simply by making changes in the parameters used or by changing the threshold values. Our main objective is to create a simple statistical analysis that is comparable in performance to the conventional approaches.

The small number of sample documents available makes the task very challenging. Thus, it is common practice to improvise and collect more information about the author's style by crawling the internet for documents that are stylistically similar to the samples. But we restrict our method to use only the samples provided in order to avoid high demand for network resources at runtime.

The effectiveness of character n-grams in stylometry has been studied extensively [5] [11]. It has been shown that character n-grams can capture valuable stylistic information which can be used to determine authorship without the need of externally sourced documents or any other information about the grammatical structure of the sample documents. To elucidate, we present the following examples. Any text with many occurrences of a question-tag like -", isn't it?" will result in a high frequency of the character 3-gram -"it?". Similarly any text written in past tense is characterized by -"ed ". Also, writings that involve a lot of questions are replete with occurrences of the -" wh" 3-gram. Thus, it is apparent that parsing the documents is not necessary to capture the necessary details about the author's preferences for various linguistic constructs, and a simpler character n-gram analysis can be used for this purpose instead. We expect this scheme to be valid for other languages as well. The optimal length of n-gram varies across languages and depends upon the morphology. This optimal length can be determined during training phase and can be passed as a parameter during the test phase.

Latent Semantic Analysis [7] is a method used to obtain a low-dimensional approximation of data represented as a matrix. This has the effect of reducing noise in the data as well as reducing the sparseness of the matrix. LSA can be accomplished by several matrix decompositions, but the Singular Value Decomposition (SVD) is the most popular method. It has been shown [8] that culling out the less significant singular values and reconstructing the matrix results in a least-squares best-fit approximation of the original matrix.

## 3 Our Approach

We generate character n-grams by sliding a window of length n along the document. These n-grams are used as the features or terms for our Term-Document matrix. This matrix is populated by the product of the local weighting and the global weighting for each term corresponding to each document. The local weighting is used to characterize the term in the current document. Several schemes like term-frequency, log-term-frequency or binary term-frequency may be used for local weighting [2] of the term. The global weighting characterizes the term across all documents in the corpus. Entropy and inverse document-frequency are used as the global weighting schemes. Different weighting schemes give different performances for different languages and

genre. Also, to reduce the effect of variability in document lengths each document row is then normalized by dividing by the square root of the document length.

We advance previous methods that apply LSA to stylometry [10] and provide a confidence measure instead of visualisations. The singular value decomposition of the matrix is obtained and the less significant singular values are culled out to obtain an approximate reconstruction of the Term-Document matrix. The document rows of the Term-document matrix are used to compute all pairwise document dissimilarities. Different dissimilarity measures based on cosine similarity or extended Jaccard similarity [13] may be used. We surmise that the authorship may be judged by studying these pairwise document dissimilarities.

We differentiate between the dissimilarities between two sample documents $\delta_{ij}$ and dissimilarities between a sample document and the disputed document $\delta_i$ . Depending on the number of sample documents available we define the following cases -

Case 1 - When there is only one sample document available, only the value $\delta_1$ can be obtained and the decision is based on this parameter only. The training corpus is used to determine a threshold value $\delta_{th1}$ such that authorship is judged to be true if $\delta_1$<$\delta_{th1}$ or false otherwise.

Case 2 - When two documents samples are available we can obtain $\delta_1$ , $\delta_2$ and $\delta_{12}$. We obtain the mean $\mu = \frac{\delta_1+\delta_2}{2}$ and $\delta_\mu = \mu - \delta_{12}$ . A threshold value $\delta_{th2}$ is obtained by training such that $\delta_\mu * \delta_{12}$<$\delta_{th2}$ implies that the given author wrote the disputed document.

Case 3 - When we have n (>2) sample documents we can obtain $\delta_{ij}$, 1 <= i, j <= n, along with n $\delta_i$ values. We define the following means $\mu_{samples} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}}{n^2}$ and $\mu_{disputed}$=$\frac{\sum_{i=1}^{n}\delta_i}{n}$ .The change in the means is measured as $\delta_\mu = \frac{\mu_{disputed}-\mu_{samples}}{\mu_{disputed}+\mu_{samples}}$. Similar computations are done for the variances of the two types of dissimilarities to obtain $\nu_{samples}$ =$\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{ij}-\mu_{samples})^2}{n^2}}$ and $\nu_{disputed}$ =$\sqrt{\frac{\sum_{i=1}^{n}(\delta_i-\mu_{disputed})^2}{n}}$. The change in the variances is measured as $\delta_\nu = \frac{\nu_{disputed}-\nu_{samples}}{\nu_{disputed}+\nu_{samples}}$.These values are used to obtain a threshold value $\delta_{th3}$ such that $\delta_\mu * \delta_\nu$<$\delta_{th3}$ is true for the cases where the disputed document is the work of the given author.

## 4 Resources

For the development of the software we use the Stylo [3] package from the CRAN repository that is distributed under the GNU GPL 3 license. Stylo provides a comprehensive collection of functions used frequently in stylometric analysis. Our software is implemented entirely in R [9] which is a popular language for statistical computing and graphics.

## 5 Result and Analysis

We trained and tested our software on the corpus provided by the PAN workshop for the Author Identification task. TIRA [4] - an automated tool for deployment and evaluation

of the software was developed at the workshop and provided as a facility to the participants. We were able to obtain competitive results while maintaining a very low runtime. The performance of our software for different categories is tabulated as follows -

**Table 1.** Performance in Author Verification Task - PAN 2014

| Corpus | AUC | C@1 | AUC * C@1 | Runtime |
|---|---|---|---|---|
| Dutch-Essays | | | | |
| Training | 0.82964 | 0.84375 | 0.70001 | 00:01:20 |
| Test | 0.65148 | 0.75 | 0.48861 | 00:01:21 |
| Dutch-Reviews | | | | |
| Training | 0.716 | 0.7373 | 0.52791 | 00:00:12 |
| Test | 0.7568 | 0.6936 | 0.52492 | 00:00:15 |
| English-Essays | | | | |
| Training | 0.7129 | 0.6969 | 0.49682 | 00:17:29 |
| Test | 0.6987 | 0.6565 | 0.4587 | 00:16:22 |
| English-Novels | | | | |
| Training | 0.8596 | 0.84 | 0.72206 | 00:22:52 |
| Test | 0.65685 | 0.57855 | 0.38002 | 02:14:27 |
| Greek-Articles | | | | |
| Training | 0.534 | 0.62 | 0.33108 | 00:12:43 |
| Test | 0.5934 | 0.6 | 0.35604 | 00:12:01 |
| Spanish-Articles | | | | |
| Training | 0.452 | 0.58 | 0.26216 | 00:08:07 |
| Test | 0.4432 | 0.56 | 0.24819 | 00:08:09 |

## 6 Conclusion and Future Work

We provide an alternative mechanism for authorship attribution with results comparable to most other approaches to the task. It paves the way for further exploration in the utility of character n-gram based analysis and effectiveness of LSA in creating a low-noise approximation of data. There is a possibility to exploit weak patterns in the corpus that follow the Zipf's Law. Also, there is a lot of scope for studying the effect of different weighting schemes, normalization methods and similarity measures on the performance of our method.

The application of Burrow's Delta [12] to stylometric applications [1] has been well studied. The Burrow's delta is Manhattan distance between document rows after the term-columns have been centered on the column mean and are normalized by dividing by the standard deviation of the column. This provides a motivation for testing other centering and normalization schemes and other distance measures like Canberra, Euclidean, etc. Several modifications [3] of the Burrow's delta have been implemented and tested and their usage could be a valuable modification in our system.

Our system performs sub-optimally for Spanish and Greek texts and this calls for a more involved study of the morphological structures of these languages to develop a more sophisticated approach.

## References


1. Burrows, J.: Delta: a measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing 17(3), 267–287 (2002), http://llc.oxfordjournals.org/content/17/3/267.abstract
2. Dumais, S.T.: Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, & Computers 23(2), 229–236 (1991), http://psychonomic.org/search/view.cgi?id=5145
3. Eder, M., Kestemont, M., Rybicki, J.: Stylometry with r: a suite of tools. In: Digital Humanities 2013: Conference Abstracts. pp. 487–89. University of Nebraska–Lincoln, Lincoln, NE (2013), http://dh2013.unl.edu/abstracts/
4. Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Recent trends in digital text forensics and its evaluation. In: Forner, P., MÃijller, H., Paredes, R., Rosso, P., Stein, B. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Springer (2013)
5. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Euzenat, J., Domingue, J. (eds.) Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Computer Science, vol. 4183, pp. 77–86. Springer Berlin Heidelberg (2006)
6. Juola, P., Stamatatos, E.: Overview of the author identification task at pan 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs (2013)
7. Landauer, T., Foltz, P., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25, 259–284 (1998)
8. Martin, D., Berry, M.: Mathematical foundations behind latent semantic analysis. Handbook of latent semantic analysis pp. 35–55 (2007)
9. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014), http://www.R-project.org/
10. Soboroff, I.M., Nicholas, C.K., Kukla, J.M., Ebert, D.S.: Visualizing document authorship using ngrams and latent semantic indexing. In: Proceedings of the 1997 Workshop on New Paradigms in information Visualization and Manipulation. pp. 43–48. Addison-Wesley, New York, NY (1997)
11. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy 21(2), 421 (March 2013)
12. Stein, S., Argamon, S.: A mathematical explanation of burrows's delta. In: Proceedings of the Digital Humanities Conference (2006)
13. Strehl, A., Ghosh, J.: Value-based customer grouping from large retail data-sets. In: Proc. SPIE Conference on Data Mining and Knowledge Discovery, Orlando. vol. 4057, pp. 33–42. SPIE (April 2000), http://strehl.com/download/strehl-spie00.pdf