Pastiche detection based on stopword rankings. Exposing impersonators of a Romanian writer

Liviu P. Dinu
Faculty of Mathematics
and Computer Science
University of Bucharest
ldinu@fmi.unibuc.ro

Vlad Niculae
Faculty of Mathematics
and Computer Science
University of Bucharest
vlad@vene.ro

Octavia-Maria Șulea
Faculty of Foreign Languages
and Literatures
Faculty of Mathematics
and Computer Science
University of Bucharest
mary.octavia@gmail.com

Abstract

We applied hierarchical clustering using Rank distance, previously used in computational stylometry, on literary texts written by Mateiu Caragiale and a number of different authors who attempted to impersonate Caragiale after his death, or simply to mimic his style. Their pastiches were consistently clustered opposite to the original work, thereby confirming the performance of the method and proposing an extension of the method from simple authorship attribution to the more complicated problem of pastiche detection.

The novelty of our work is the use of frequency rankings of stopwords as features, showing that this idea yields good results for pastiche detection.

1 Introduction

The postulated existence of the human stylome has been thoroughly studied with encouraging results. The term *stylome*, which is currently not in any English dictionaries, was recently defined as a linguistic fingerprint which can be measured, is largely unconscious, and is constant (van Halteren et al., 2005).

Closely related to the problem of authorship attribution lies the pastiche detection problem, where the fundamental question is: Can the human stylome be faked in order to trick authorship attribution methods? There are situations where certain authors or journalists have tried to pass their own work as written by someone else. A similar application is in forensics, where an impersonator is writing letters or messages and signing with someone else's name, especially online.

It is important to note that sometimes pastiches are not intended to deceive, but simply as an exercise in mimicking another's style. Even in this case, the best confirmation that the author of the pastiche can get is if he manages to fool an authorship attribution algorithm, even if the ground truth is known and there is no real question about it.

Marcus (1989) identifies the following four situation in which text authorship is disputed:

- A text attributed to one author seems non-homogeneous, lacking unity, which raises the suspicion that there may be more than one author. If the text was originally attributed to one author, one must establish which fragments, if any, do not belong to him, and who are their real authors.
- A text is anonymous. If the author of a text is unknown, then based on the location, time frame and cultural context, we can conjecture who the author may be and test this hypothesis.
- If based on certain circumstances, arising from literature history, the paternity is disputed between two possibilities, A and B, we have to decide if A is preferred to B, or the other way around.
- Based on literary history information, a text seems to be the result of the collaboration of two authors, an ulterior analysis should establish, for each of the two authors, their corresponding text fragments.

We situate ourselves in a case similar to the third, but instead of having to choose between two authors, we are asking whether a group of texts were indeed written by the claimed author or by someone else. Ideally, we would take samples authored by every possible impersonator and run a

Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection, pages 72–77,
Avignon, France, April 23 - 27 2012.

multi-class classifier in order to estimate the probability that the disputed work is written by them or by the asserted author. Such a method can give results if we know who the impersonator can be, but most of the time that information is not available, or the number of possible impersonators is intractably large.

In the case of only one impersonator, the problem can simply be stated as authorship attribution with a positive or a negative answer. However, when there are a number of people separately writing pastiches of one victim's style, the extra information can prove beneficial in an unsupervised learning sense. In this paper we analyze the structure induced by the Rank Distance metric using frequencies of stopwords as features, previously applied for authorship attribution, on such a sample space. The assumption is that trying to fake someone else's stylome will induce some consistent bias so that new impersonators can be caught using features from other pastiche authors.

2 The successors of Mateiu Caragiale

Mateiu Caragiale, one of the most important Romanian novelists, died in 1936, at the age of 51, leaving behind an unfinished novel, *Sub pecetea tainei*. Some decades later, in the 70's, a rumor agitated the Romanian literary world: it seemed that the ending of the novel had been found. A few human experts agreed that the manuscript is in concordance with Mateiu's style, and in the next months almost everybody talked about the huge finding. However, it was suspicious that the writer who claimed the discovery, Radu Albala, was considered by the critics to be one of the closest stylistic followers of Mateiu Caragiale. When the discussions regarding the mysterious finding reached a critical mass, Albala publically put a stop to them, by admitting that he himself had written the ending as a challenge - he wanted to see how well he could deceive the public into thinking the text in question was written by Mateiu himself.

Other authors attempted to write different endings to the novel, but without claiming Caragiale's paternity, like Albala did. Around the same time, Eugen Bălan also set to continue the unfinished novel, as a stylistic exercise. He addressed a separate storyline than Albala's. Later, Alexandru George also attempted to finish the novel, claiming that his ending is the best. Unfortunately there is only one copy of George's work, and we couldn't obtain it for this study.

In 2008, Ion Iovan published the so-called *Last Notes of Mateiu Caragiale*, composed of sections written from Iovan's voice, and another section in the style of a personal diary describing the life of Mateiu Caragiale, suggesting that this is really Caragiale's diary. This was further strengthened by the fact that a lot of phrases from the diary were copied word for word from Mateiu Caragiale's novels, therefore pushing the style towards Caragiale's. However, this was completely a work of fiction, the diary having been admittedly imagined and written by Iovan.

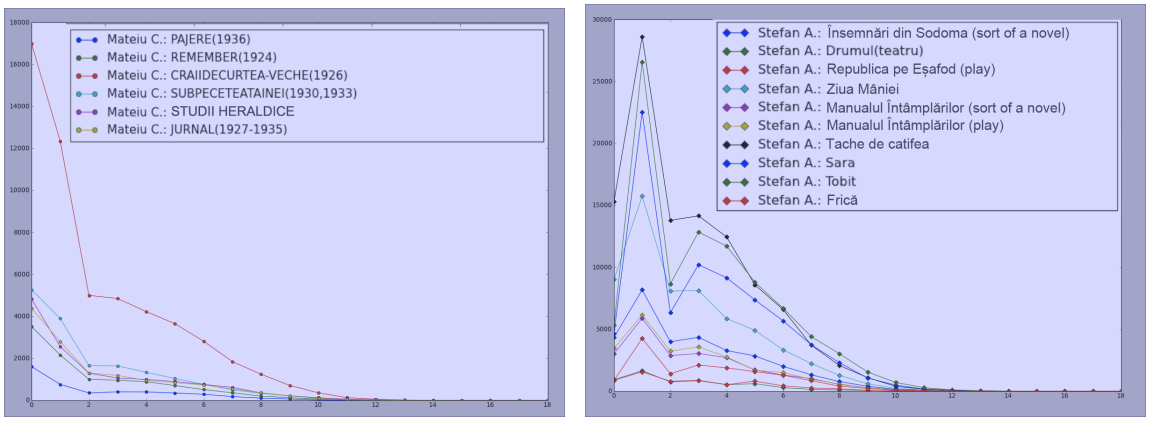
Another noteworthy case is the author Ştefan Agopian. He never attempted to continue Mateiu Caragiale's novel, but critics consider him one of his closest stylistic successors. Even though not really a pastiche, we considered worth investigating how such a successor relates to the impersonators.

3 Simple visual comparisons

The pioneering methods of Mendenhall (Mendenhall, 1901) on the subject of authorship attribution, even though obsolete by today's standards, can be used to quickly examine at a glance the differences between the authors, from certain points of view. The Mendenhall plot, showing frequency versus word length, does not give an objective criterion to attribute authorship, but as an easy to calculate statistic, it can motivate further research on a specific attribution problem.

A further critique to Mendenhall's method is that different distributions of word length are not necessary caused by individual stylome but rather by the genre or the theme of the work. This can further lead to noisy distributions in case of versatile authors, whereas the stylome is supposed to be stable.

Even so, the fact that Mateiu Caragiale's Mendenhall distribution has its modes consistently in a different position than the others, suggests that the styles are different, but it appears that Caragiale's successors have somewhat similar distributions. This can be seen in figure 3. In order to evaluate the questions *How different, how similar?*, and to make a more objective judgement on authorship attribution, we resort to pairwise distance-based methods.



(a) Mateiu Caragiale

(b) Ştefan Agopian

(c) Radu Albala

(d) Ion Iovan

Figure 1: Mendenhall plots: frequency distribution of word lengths, showing similarities between the other authors, but differences between them and Mateiu Caragiale.

şi în să se cu o la nu a ce mai din pe un că ca mă fi care era lui fără ne pentru el ar dar îl tot am mi însă într cum când toate al aa după până decât ei nici numai dacă eu avea fost le sau spre unde unei atunci mea prin ai atât au chiar cine iar noi sunt acum ale are asta cel fie fiind peste această a cele face fiecare nimeni încă între aceasta aceea acest acesta acestei avut ceea cât da făcut noastră poate acestui alte celor cineva către lor unui altă ați dintre doar foarte unor vă aceste astfel avem aveți cei ci deci este suntem va vom vor de

Table 1: The 120 stopwords extracted as the most frequent words in the corpus.

In order to speak of distances, we need to represent the samples (the novels) as points in a metric space. Using the idea that stopword frequencies are a significant component of the stylome, and one that is difficult to fake (Chung and Pennebaker, 2007), we first represented each work as a vector of stopword frequencies, where the stopwords are chosen to be the most frequent words from all the concatenated documents. The stopwords can be seen in table 1. Another useful visualisation method is the Principal Components Analysis, which gives us a projection from a high-dimensional space into a low-dimensional one, in this case in 2D. Using this stopword frequency representation, the first principal components plane looks like figure 3.

4 Distances and clustering

In (Popescu and Dinu, 2008), the use of rankings instead of frequencies is proposed as a smoothing method and it is shown to give good results for computational stylometry. A ranking is simply an ordering of items; in this case, the representation of each document is the ranking of the stopwords in that particular document. The fact that a specific function word has the rank 2 (is the second most frequent word) in one text and has the rank 4 (is the fourth most frequent word) in another text can be more directly relevant than the fact that the respective word appears 349 times in the first text and only 299 times in the second.

Rank distance (Dinu, 2003) is an ordinal metric able to compare different rankings of a set of objects. In the general case, Rank distance works for

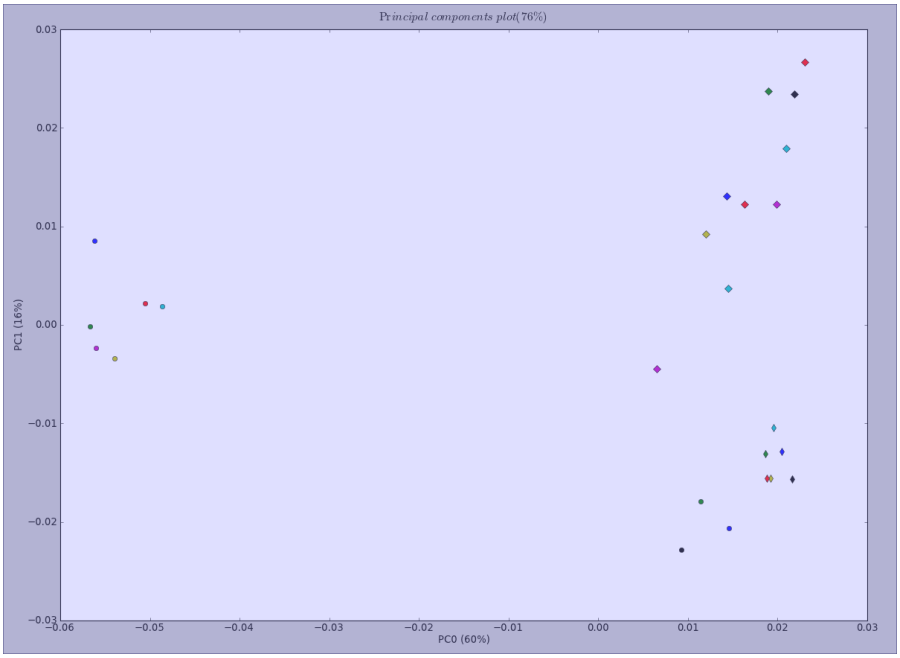


Figure 2: Principal components plot. Works are colour coded like in figure 3. The cluster on the left consists only of novels by Mateiu Caragiale. Individual authors seem to form subclusters in the right cluster.

rankings where the support set is different (for example, if a stopword would completely be missing from a text). When this is not the case, we have the following useful property:

A ranking of a set of n objects is a mapping $\sigma : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., n\}$ where $\sigma(i)$ will represent the place (rank) of the object indexed as i such that if $\sigma(q) < \sigma(p)$ word q is more frequent than word p. The Rank distance in this case is simply the distance induced by L_1 norm on the space of vector representations of permutations:

$$D(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)| \qquad (1)$$

This is a distance between what is called full rankings. However, in real situations, the problem of *tying* arises, when two or more objects claim the same rank (are ranked equally). For example, two or more function words can have the same frequency in a text and any ordering of them would be arbitrary.

The Rank distance allocates to tied objects a number which is the average of the ranks the tied objects share. For instance, if two objects claim the rank 2, then they will share the ranks 2 and 3 and both will receive the rank number $(2+3)/2 = 2.5$. In general, if k objects will claim the same rank and the first x ranks are already used by other objects, then they will share the ranks $x + 1, x + 2, \ldots, x + k$ and all of them will receive as rank the number: $\frac{(x+1)+(x+2)+...+(x+k)}{k} = x + \frac{k+1}{2}$. In this case, a ranking will be no longer a permutation ($\sigma(i)$ can be a non integer value), but the formula (1) will remain a distance (Dinu, 2003).

Even though computationally the formula (1) allows us to use the L_1 distance we will continue using the phrase Rank distance to refer to it, in order to emphasize that we are measuring distances between rankings of stopwords, not L_1 distances between frequency values or anything like that.

Hierarchical clustering (Duda et al., 2001) is a bottom-up clustering method that starts with the most specific cluster arrangement (one cluster for each sample) and keeps joining the *nearest* clusters, eventually stopping when reaching either a stopping condition or the most general cluster arrangement possible (one cluster containing all the samples). When joining two clusters, there are many possible ways to specify the distance between them. We used *complete linkage*: the distance between the most dissimilar points from the two clusters. The resulting clustering path, visualised a dendrogram, is shown in figure 4.

The use of clustering techniques in authorship attribution problems has been shown useful by Labbé and Labbé (2006); Luyckx et al. (2006). Hierarchical clustering with Euclidean distances

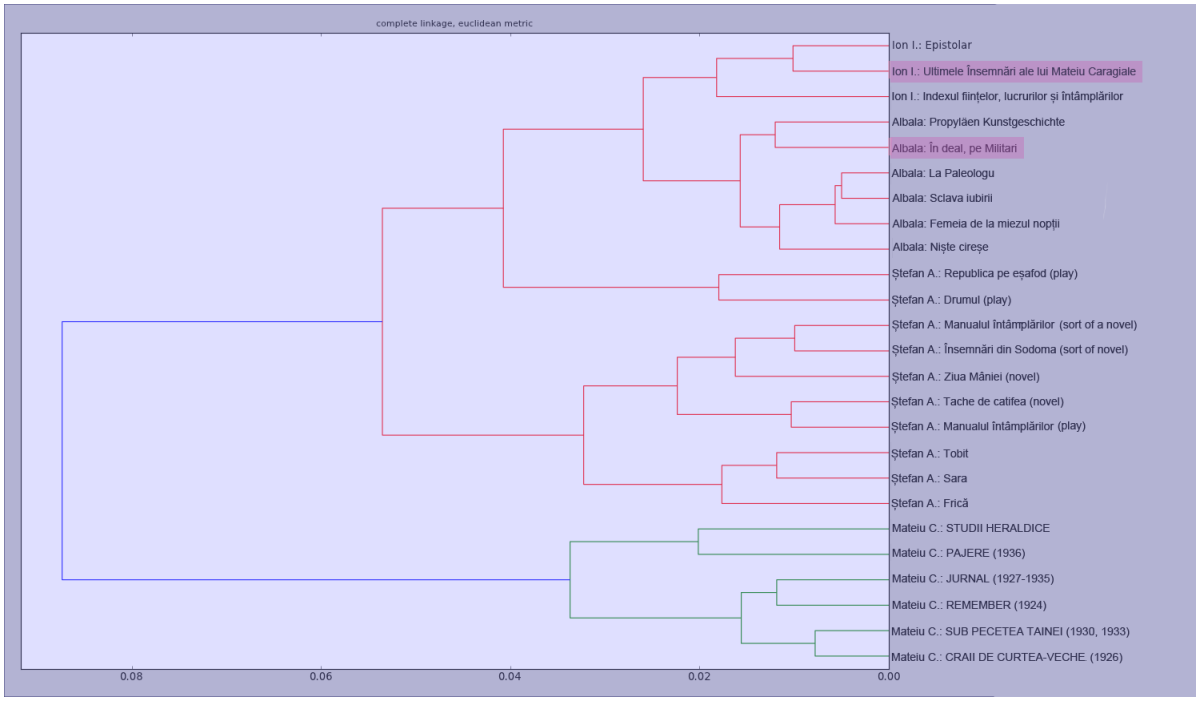


Figure 3: Dendrogram showing the results of hierarchical clustering using the L_2 (euclidean) distance.

has been used for pastiche detection in (Somers and Tweedie, 2003). The novelty of our work is the use of rankings as features, and using the L_1 distance (equivalent to the Rank distance for this particular case). (Somers and Tweedie, 2003) shows how the Euclidean distance clusters mostly works by the same author at the finest level, with a few exceptions. On the data from our problem, we observed a similar problem. The Euclidean distance behaves in a less than ideal fasion, joining some of Agopian's works with the cluster formed by the other authors (see figure 3), whereas the Rank distance always finds works by the same author the most similar at the leaves level (with the obvious exception of Eugen Bălan's text, because it is his only available text).

Reading the dendrogram in the reverse order (top to bottom), we see that for $k = 2$ clusters, one corresponds to Mateiu Caragiale and the other to all of his successors. In a little finer-grained spot, there is a clear cluster of Ştefan Agopian's work, the (single) text by Eugen Bălan, and a joint cluster with Radu Albala and Ion Iovan, which also quickly breaks down into the separate authors. The fact that there is no k for which all authors are clearly separated in clusters can be attributed to the large stylistic variance exhibited by Ştefan Agopian and Mateiu Caragiale, whose clusters break down more quickly.

These results confirm our intuition that rankings of stopwords are more relevant than frequencies, when an appropriate metric is used. Rank distance is well-suited to this task. This leads us to believe that if we go back and apply our methods to the texts studies in (Somers and Tweedie, 2003), an improvement will be seen, and we intend to further look into this.

5 Conclusions

We reiterate that all of the authors used in the study are considered stylistically similar to Mateiu Caragiale by the critics. Some of their works, highlighted on the graph, were either attributed to Caragiale (by Albala and Iovan), or intended as pastiche works continuing Caragiale's unfinished novel.

A key result is that with this models, all of these successors prove to be closer to each other than to Mateiu Caragiale. Therefore, when faced with a new problem, we don't have to seed the system with many works from the possible authors (note that we used a single text by Bălan): it suffices to use as seeds texts by one or more authors who are stylistically and culturally close to the claimed author (in this case, Mateiu Caragiale). Clustering with an appropriate distance such as Rank dis-

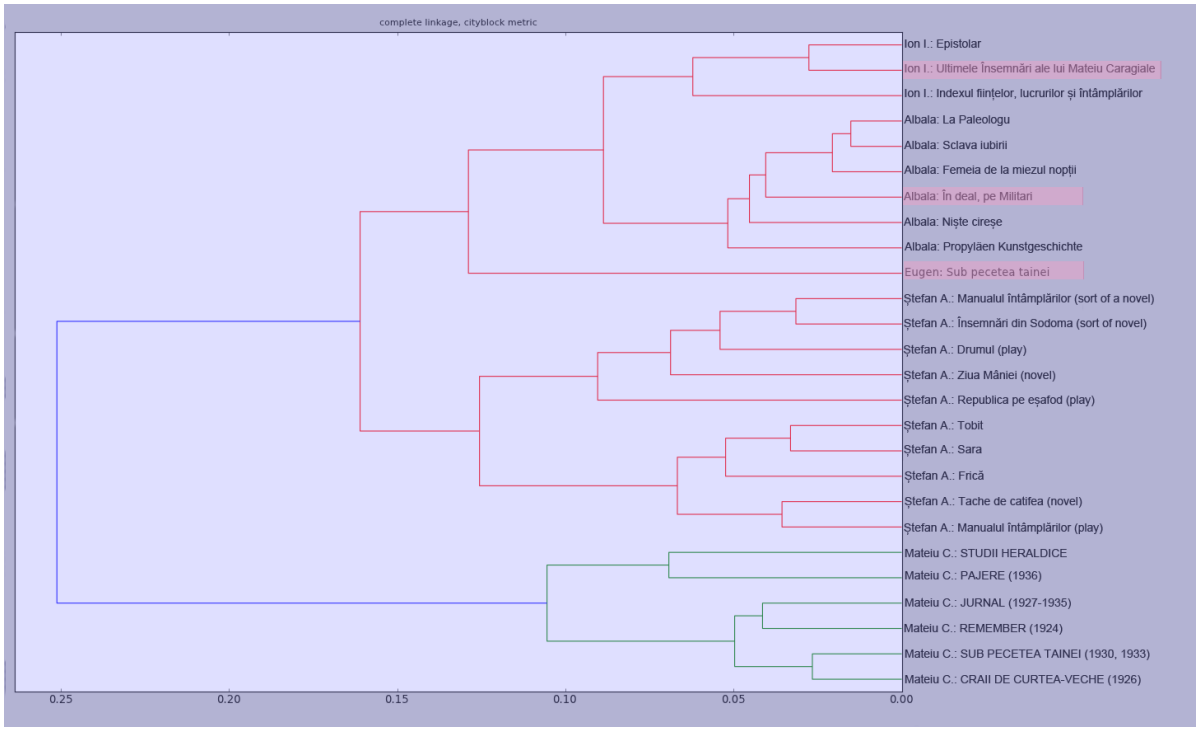


Figure 4: Dendrogram showing the results of hierarchical clustering using L_1 distance on stopword rankings (equivalent to Rank distance).

tance will unmask the pastiche.

References

Cindy Chung and James Pennebaker. The psychological functions of function words. *Social communication: Frontiers of social psychology*, pages 343–359, 2007.

Liviu Petrisor Dinu. On the classification and aggregation of hierarchies with different constitutive elements. *Fundamenta Informaticae*, 55 (1):39–50, 2003.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd ed.)*. Wiley-Interscience Publication, 2001.

Cyril Labbé and Dominique Labbé. A tool for literary studies: Intertextual distance and tree classification. *Literary and Linguistic Computing*, 21(3):311–326, 2006.

Kim Luyckx, Walter Daelemans, and Edward Vanhoutte. Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of LREC-2006, the fifth International Language Resources and Evaluation Conference*, pages 30–35, 2006.

Solomon Marcus. *Inventie si descoperire*. Ed. Cartea Romaneasca, Bucuresti, 1989.

T C Mendenhall. A mechanical solution of a literary problem. *Popular Science Monthly*, 60(2): 97–105, 1901.

Marius Popescu and Liviu Petrisor Dinu. Rank distance as a stylistic similarity. In *COLING (Posters)'08*, pages 91–94, 2008.

Harold Somers and Fiona Tweedie. Authorship attribution and pastiche. *Computers and the Humanities*, 37:407–429, 2003. ISSN 0010-4817. 10.1023/A:1025786724466.

Hans van Halteren, R. Harald Baayen, Fiona J. Tweedie, Marco Haverkort, and Anneke Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, pages 65–77, 2005.