UniNE at CLEF 2017: Author Clustering

Notebook for PAN at CLEF 2017

Mirco Kocher and Jacques Savoy

Computer Science Dept., University of Neuchâtel, Switzerland
{Mirco.Kocher, Jacques.Savoy}@unine.ch

Abstract. This paper describes and evaluates an effective unsupervised author clustering and authorship linking model called SPATIUM. The suggested strategy can be adapted without any difficulty to different languages (such as Dutch, English, and Greek) in different text genres (e.g., newspaper articles and reviews). As features, we suggest using the *m* most frequent terms (isolated words and punctuation symbols) or the *m* most frequent character *n*-grams of each text. Applying a simple distance measure, we determine whether there is enough indication that two texts were written by the same author. The evaluations are based on 60 training and 120 test problems (PAN AUTHOR CLUSTERING task at CLEF 2017). Using the most frequent terms results in a higher clustering precision, while using the most frequent character *n*-grams of letters gives a higher clustering recall. An analysis to assess the variability of the performance measures indicates that we have a system working stable independent of the underlying text collection and that our parameter choices did not over-fit to the training data.

1 Introduction

The authorship attribution problem is an interesting problem in computational linguistics but also in applied areas such as criminal investigation and historical studies where knowing the author of a document (such as a ransom note) may be able to save lives [14]. With the Web 2.0 technologies, the number of anonymous or pseudonymous texts is increasing and in many cases one person writes in different places about different topics (e.g., multiple blog posts written by the same author). Therefore, proposing an effective algorithm to the authorship problem presents a real interest. In this case, the system must regroup all texts by the same author (possibly written about different text topics) into the same group or cluster. A justification supporting the proposed answer and a probability that the given answer is correct can be given to improve the confidence attached to the response [10].

This author clustering task is more demanding than the classical authorship attribution problem. Given a document collection the task is to group documents written by the same author such that each cluster corresponds to a different author. The number of distinct authors whose documents are included is not given. For example, based on a set of passages extracted from larger documents, we should first determine the number of authors *k* and then regroup the texts into *k* clusters according to their real author. This task can also be viewed as establishing authorship links between texts and is related to the PAN 2015 task of authorship verification.

This paper is organized as follows. The next section presents the test collections and the evaluation methodology used in the experiments. The third section explains our proposed algorithm called SPATIUM. Then, we evaluate the proposed scheme on 60 training problems and compare it to the best performing schemes using 120 different test problems. The last section explains our parameter choices and provides a sensibility assessment. A conclusion draws the main findings of this study.

2 Test Collections and Evaluation Methodology

The evaluation was performed using the *TIRA* platform, which is an automated tool for deployment and evaluation of the software [3]. The data access is restricted such that during a software run the system is encapsulated and thus ensuring that there is no data leakage back to the task participants [9]. This evaluation procedure also offers a fair evaluation of the time needed to produce an answer.

During the PAN CLEF 2017 evaluation campaign, six corpora (or test collections) were built each containing 30 problems (10 for training and 20 for testing). In each problem, all the texts matched the same language, are in the same text genre, and are single-authored, but they may differ in text-length and can be cross-topic [14]. The number of distinct authors is not given. In this context, the task is defined as:

> *Given a problem of up to 50 short documents, identify*
> *authorship links and groups of documents by the same author.*

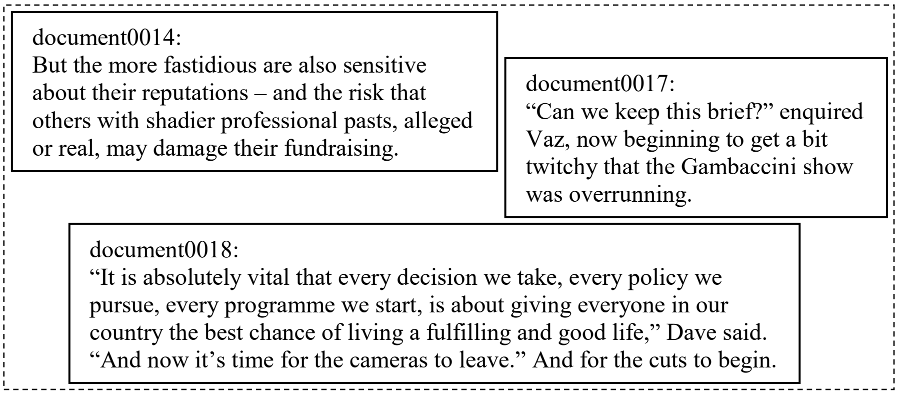
The six corpora are a combination of one of three languages (English, Dutch, or Greek) and one of two genres (newspaper articles or reviews). An overview of these corpora is depicted in Table 1. Considering the six benchmarks we have 120 problems to test and 60 problems to train (pre-evaluate) our system. The training set was used to evaluate our approach and the test set was used to compare our results with other participants of the PAN CLEF 2017 campaign. This year, everyone had access to the test data twice. This means we can train and test a basic approach, improve it or provide a different approach, and then test it again for the second and final run.

Table 1. PAN CLEF 2017 *training* corpora statistics.

Corpus	Training Sets			
	Texts	Authors	Single	Terms
English Newspaper (EN)	20.0	5.6 [3-10]	1.8 [0-6]	62 [56-67]
English Reviews (ER)	19.4	6.1 [4-10]	1.9 [0-5]	73 [70-77]
Dutch Newspaper (DN)	20.0	5.3 [4-8]	2.0 [0-5]	59 [53-66]
Dutch Reviews (DR)	18.2	6.5 [5-8]	0.3 [0-2]	159 [143-186]
Greek Newspaper (GN)	20.0	6.0 [4-8]	1.5 [0-5]	76 [66-88]
Greek Reviews (GR)	20.0	6.1 [4-9]	2.1 [0-6]	62 [53-70]

For each corpus, we have 10 problems in the training dataset containing the average number of texts as given under the label "Texts". The number of distinct authors on average together with the range for each corpus is indicated in the column "Authors", and the average with the minimum and maximum number of authors with only a single document is presented under the label “Single”. Finally, the average number of terms (isolated words and punctuation symbols) is given in the column “Terms”. For example, with the English newspaper collection (training set), 20 texts are written, on average, by 5.6 authors and we can find 1.8 authors who wrote only one single article. These metrics are not available for the test corpora because the datasets remain undisclosed thanks to the *TIRA* system. We only know that the same combinations of language and genre are present.

In Table 1 we see that the number of words is rather small. In Figure 1 we show three texts extracted from a problem containing articles written in the English language. The represented texts are the full unmodified documents as available in problem001. Notice that document0014 and document0017 are a single sentence, and the latter is so short that it would fit in a single Twitter tweet (it contains less than 140 characters). When analyzing the texts, we should detect a shared authorship between document0017 and document0018, but not with document0014 as this was written by someone else. The limited length of those documents is the main difficulty of this year’s author clustering task.



document0014:
But the more fastidious are also sensitive about their reputations – and the risk that others with shadier professional pasts, alleged or real, may damage their fundraising.

document0017:
“Can we keep this brief?” enquired Vaz, now beginning to get a bit twitchy that the Gambaccini show was overrunning.

document0018:
“It is absolutely vital that every decision we take, every policy we pursue, every programme we start, is about giving everyone in our country the best chance of living a fulfilling and good life,” Dave said. “And now it’s time for the cameras to leave.” And for the cuts to begin.

Figure 1. Sample texts from problem001.

When inspecting the training corpora, the number of words available is rather small (overall in mean 82 terms for each text). Since there are some authors who only wrote a single text we should only cluster two texts if there are enough evidences for a single authorship.

During the PAN CLEF 2017 campaign, a system must return two outputs in a JSON structure for each problem. First, the detected groups should be written to a file indicating the author clustering. Each text must belong to exactly one cluster; thus, the clusters must be non-overlapping. Second, a list of text pairs with a probability of having the same author should be written to another file representing the authorship links.

As performance measure, two evaluation measures were used during the PAN CLEF campaign. The first performance measure is the BCubed F_1 to evaluate the clustering output [1]. This value is the harmonic mean of the precision and recall associated to each document. The document precision represents how many documents in the same cluster are written by the same author and therefore measures the purity of its cluster. Symmetrically, the recall associated to one document represents how many documents from that author appear in its cluster and therefore measures the completeness of its cluster.

As another measure, the PAN CLEF campaign adopts the mean average precision (MAP) measure for the authorship links between document pairs [8]. This evaluation measure provides a single-figure measure of quality across recall levels. The MAP is roughly the average area under the precision-recall curve for a set of problems. Therefore, this measure gives more emphasis on the first positions and a misclassification with a lower probability is less penalized. MAP does not punish verbosity, i.e., every true link counts even when appearing near the end of the ranked list. Therefore, by providing all possible authorship links, one can attempt to maximize MAP [13].

3 Simple Clustering Algorithm

We suggest an unsupervised approach based on a simple feature extraction and distance measure called SPATIUM (Latin word meaning distance). The selected stylistic features correspond to the top *m* most frequent terms (isolated words without stemming but with the punctuation symbols) in the first run as in the last year [5] and additionally the *m* most frequent character *n*-grams for the second run. The features are selected solely based on the frequency in the query text. For determining the value of *m*, previous studies have shown that a value between 200 and 300 tends to provide the best performance [2, 10]. The texts were only paragraphs so the effective number of features *m* was set to at most 200 but was in most cases well below. The length of the *n*-grams was set to *n=6* characters to ease the analysis of the most pertinent features. Unlike in the previous year [5], we did not remove the words appearing only once (*hapax legomenon*) in the text due to the limited size of each document (see Table 1). For instance, in document0017 depicted in Figure 1 every term would be deleted if the *hapax legomenon* would be ignored.

To measure the distance between a Test A and another Text B, SPATIUM uses a variant of the L^1-norm called Canberra. This distance suggests that the absolute differences of the individual features are normalized based on the sum of them as indicated in Equation 1.

$$\Delta(A, B) = \Delta_{AB} = \sum_{i=1}^{m} \frac{|P_A[f_i] - P_B[f_i]|}{P_A[f_i] + P_B[f_i]} \qquad (1)$$

where *m* indicates the number of features (words and punctuation symbols, or character *n*-grams), and $P_A[f_i]$ and $P_B[f_i]$ represent the estimated occurrence probability of the feature f_i in the first Text A and in the other Text B respectively. To estimate these probabilities, we divide the feature occurrence frequency (ff_i) by the sum of all features of the corresponding text (*n*), Prob[f_i] = ff_i / *n*, without smoothing and therefore accepting a probability of 0.0 in Text B. This distance measure is not symmetric due to the choice of the features to be include in the computation.

Observing a small value for Δ_{AB} provides evidence that both documents are written by the same author. On the other hand, a large value suggests the opposite. The real problem consists in defining precisely what a “small distance value” is. To verify whether the resulting Δ_{AB} value is small or rather large, a comparison basis must be determined.

To achieve this with a specific problem, the distance *from* Text A to all other texts is computed (or $\Delta(A,j)$). From this distribution, the mean (denoted $m(A,.)$) and standard deviation ($std(A,.)$) are estimated. Moreover, the distribution of distance values *to* Text B (or $\Delta(j,B)$) can be computed to provide the mean $m(.,B)$ and the standard deviation $std(.,B)$ of the intertextual distances *to* Text B.

As a first definition of a "small" distance, we can assume that a small distance value *from* Text A must respect Eq. 2. In this formulation, δ is a parameter to be fixed.

$$Hint\ 1\text{:}\quad \Delta(A,j) \leq \phi(A,.) = m(A,.) - \delta * std(A,.) \tag{2}$$

Similarly, a small distance *to* Text B can be defined as:

$$Hint\ 2\text{:}\quad \Delta(j,B) \leq \phi(.,\mathrm{B}) = m(.,\mathrm{B}) - \delta * std(.,\mathrm{B}) \tag{3}$$

With these two decision rules, one can verify if a distance Δ_{AB} is small in comparison with all distances *from* Text A (Eq. 2) or all distances *to* Text B (Eq. 3). In the same way, one can verify whether the resulting Δ_{BA} value is small or rather large. Therefore, we propose to create two additional decision rules with Eq. 4 (based on the distribution of distance values *from* Text B) and Eq. 5 (for distance *to* Text A) as follows:

$$Hint\ 3\text{:}\quad \Delta(B,j) \leq \phi(B,.) = m(B,.) - \delta * std(B,.) \tag{4}$$

$$Hint\ 4\text{:}\quad \Delta(j,A) \leq \phi(.,\mathrm{A}) = m(.,\mathrm{A}) - \delta * std(.,\mathrm{A}) \tag{5}$$

An authorship between Text A and Text B is expected if at least two of the four hints are satisfied. For the clustering output, we use the single linkage strategy. For the list of links, we must rank each pair of texts by the certainty that they have a shared authorship. To determine the probability of a correct author linking we include both the number of satisfied hints *h* and the absolute distance between two texts in the computation [6]. A link with *h* hints fulfilled gets a probability between $h/5$ and $(h+1)/5$, where the final score depends on the other text pairs that also satisfy *h* hints.

4 Evaluation

Since our system is based on an unsupervised approach we could directly evaluate it using the training set. In Table 2a, we have reported the same performance measure applied during the PAN CLEF campaign, namely the BCubed F_1 (with the clustering precision and recall) and the AP using the most frequent terms from our first run and in Table 2b with the most frequent character *6*-grams as used in the second run. Each corpus consists of 10 problems and we report the average of them in the last row. The final score is the arithmetic mean between the BCubed F_1 and the MAP.

The algorithm returns similar results over all corpora and seems to work stable independent of the text genre and language. But we can see that from the first to the second approach (from Table 2a to Table 2b) that the precision drops significantly and the recall increases notably. Overall, the approach with *6*-grams results in a slightly higher performance of the clustering output (BCubed F_1), the authorship linking (MAP), and the Final score (+2.4% difference, +5.3% change).

Table 2a. Evaluation for the *training* corpora using the most frequent *terms* (first run).

Corpus	**Final**	F_1	Precision	Recall	MAP
EN	0.4432	0.4836	0.8351	0.3533	0.4029
ER	0.4604	0.5332	0.8599	0.4098	0.3876
DN	0.4635	0.4762	0.8905	0.3344	0.4508
DR	0.4649	0.5988	0.9247	0.4464	0.3310
GN	0.4362	0.5316	0.8630	0.3863	0.3409
GR	0.4193	0.4929	0.8725	0.3515	0.3458
Overall	0.4479	0.5194	0.8743	0.3803	0.3765

Table 2b. Evaluation for the *training* corpora using the most frequent *6-grams* (second run).

Corpus	**Final**	F_1	Precision	Recall	MAP
EN	0.4700	0.5338	0.7328	0.4423	0.4063
ER	0.5171	0.5948	0.6649	0.6356	0.4394
DN	0.5321	0.5700	0.7844	0.4783	0.4943
DR	0.4299	0.5476	0.5802	0.5558	0.3122
GN	0.4551	0.5491	0.7430	0.4673	0.3611
GR	0.4265	0.5388	0.7241	0.4625	0.3142
Overall	0.4718	0.5557	0.7049	0.5070	0.3879

The test set is then used to rank the performance of all 6 participants in this task. Based on the same evaluation methodology, we achieve the results depicted in Table 3a and Table 3b corresponding to the six test corpora.

Table 3a. Evaluation for the *test* corpora using the most frequent *terms* (first run).

Corpus	**Final**	F_1	Precision	Recall	MAP
EN	0.4776	0.4923	0.8860	0.3498	0.4628
ER	0.4320	0.5315	0.8089	0.4052	0.3325
DN	0.4537	0.5023	0.8779	0.3590	0.4051
DR	0.4575	0.6012	0.8973	0.4606	0.3138
GN	0.4339	0.4551	0.8763	0.3190	0.4127
GR	0.4255	0.4930	0.8925	0.3501	0.3581
Overall	0.4467	0.5126	0.8732	0.3740	0.3808

Table 3b. Evaluation for the *test* corpora using the most frequent *6-grams* (second run).

Corpus	**Final**	F_1	Precision	Recall	MAP
EN	0.5384	0.6068	0.7539	0.5244	0.4700
ER	0.4777	0.5696	0.6609	0.5731	0.3859
DN	0.5130	0.5860	0.7381	0.5291	0.4399
DR	0.4209	0.5349	0.5428	0.5597	0.3068
GN	0.4779	0.5107	0.7520	0.4214	0.4451
GR	0.4123	0.5021	0.6162	0.4876	0.3225
Overall	0.4734	0.5517	0.6773	0.5159	0.3951

As we can see, the final scores with all corpora are as expected from the training set with both approaches. We see a very similar performance when comparing it with the training set. Therefore, the system seems to perform stable independent of the underlying text collection and is not over-fitted to the data.

To put those values in perspective we can see in Table 4 our result in comparison with the other participants using macro-averaging for the effectiveness measures and showing the total runtime sorted by the final score. Overall, we are ranked 2nd out of 6 approaches.

Table 4. Evaluation comparison.

Rank	Participants	**Final**	F_1	MAP	Runtime (h:m:s)
1	Gómez-Adorno et al.	**0.5142**	0.5732	0.4552	00:02:05
2	Kocher & Savoy	**0.4734**	0.5517	0.3951	00:00:41
3	García et al.	**0.4724**	0.5647	0.3800	00:15:49
4	Halvani & Graner	**0.3441**	0.5488	0.1394	00:12:25
5	Karaś	**0.2958**	0.4663	0.1252	00:00:26
6	Alberts	**0.2846**	0.5276	0.0416	00:01:45

Generally, there are only small differences in the BCubed F_1 between the participants. Conversely, the MAP shows substantial variations and impacts the final score the most. The runtime only shows the actual time spent to classify the test set. On *TIRA*[1] there was the possibility to first train the system using the training set which had no influence on the final runtime. Since we have an unsupervised system it did not need to train any parameters, but this possibility might have been used by other participants.

Overall, we achieve excellent results using a rather simple and fast approach in comparison with the other solutions.

In text categorization studies, we are convinced that a deeper analysis of the evaluation results is important to obtain a better understanding of the advantages and drawbacks of a suggested scheme. By just focusing on overall performance measures, we only observe a general behavior or trend without being able to acquire a better explanation of the proposed assignment. To achieve this deeper understanding, we could analyze some problems extracted from the English corpus. The relative frequency (or probability) differences with very frequent tokens such as *the*, *(comma)*, *to*, or *and* can explain the decision. The confirmation of an authorship link is in many cases based on topical words and names that two texts share, like *labour*, *party*, *people*, *Cameron*, or *work*.

5 Parameter Choices

Our approach uses a few parameters to solve the clustering task. The main influences on the performance are the choice of the distance measure, the threshold value δ, and the feature selection scheme. Taking a decision solely on the outcome in the training data could lead to over-fitting. A leaving-one-out or a fold cross-validation is not possible in this task. Instead the bootstrap approach can be used. In this perspective, for each problem, the system must generate S new random bootstrap samples. More precisely, for each text, we will create $S = 200$ new copies having the same text length.

[1] http://www.tira.io/task/author-clustering/

For each copy the probability of choosing one given feature (word and punctuation symbol, or *n*-gram) depends on its relative frequency in the original text. This drawing is done with replacement; thus, the underlying probabilities are stable. Each resulting text must be viewed as a bag-of-words. As the syntax is not respected, each bootstrap text is not readable but still reflects the stylistic aspects as analyzed by the SPATIUM approach.

For each of the original 60 training problems (Table 1) we now have 200 generated problems of bootstrap samples and can compare different parameter choices. In Table 5 we analyze several distance measures and report the mean of the Final score achieved with the 200*60 new problems together with the limit of ±2 standard deviations σ corresponding to a confidence interval of 95.4%. Furthermore, the last two columns show the mean of the BCubed F₁ and MAP over the 200 bootstrap samples and 60 problems.

Table 5. Results of various distance measures after applying the bootstrap estimation.

Distance	**Final**			BCubed F_1	MAP
	$\bar{x}$	$\bar{x} - 2\sigma$	$\bar{x} + 2\sigma$		
Manhattan	0.3933	0.3105	0.4761	0.4848	0.3018
Euclidean	0.3766	0.3158	0.4374	0.4671	0.2862
Canberra	0.4142	0.3387	0.4898	0.4977	0.3308
Clark	0.4199	0.3421	0.4976	0.5074	0.3324
Matusita	0.4156	0.3284	0.5028	0.5021	0.3291
KLD	0.4145	0.3294	0.4997	0.5001	0.3289
Cosine	0.3881	0.3069	0.4694	0.4851	0.2911
Dice	0.3974	0.3080	0.4867	0.4904	0.3043

In a previous study [7] we found that Canberra and Clark work better on average in author profiling tasks than Cosine and Euclidean. We can again see the same distinction for this clustering task. In Table 5 we can also see that there is no significant difference between Canberra, Clark, Matusita, and KLD. For instance, between Canberra and Clark we only observe a relative change of +1.4% in the mean final score with the bootstrap approach, which isn't a substantial improvement that justifies changing our model.

The next parameter to optimize is the threshold value δ that indicates the willingness of having more or less strict assignments. A smaller value for δ generates more potential links between texts and thus increases the risk of observing incorrect assignments. For a Gaussian distribution, common choices are $\delta = 1.96$ to take account of 95%, $\delta = 1.64$ which contains 90%, $\delta = 1.28$ to include 80%, and $\delta = 1.0$ to take in 66.3%. If a corpus is composed of many authors with each cluster contains only a few items, the parameter δ should be fixed at a relatively higher level. In our system from 2016, we set $\delta = 2.0$ because of the small average cluster size [5]. With the dataset from 2017 the number of authors with only a single text is lower and there are more grouped up documents. Therefore, we decreased the threshold parameter in the current system to $\delta = 1.64$.

Figure 2 shows the mean of the BCubed F₁, the MAP, and the Final score for different δ values when using the bootstrap approach. We can see that there was a lot of potential to improve the clustering outcome (highest line on top, BCubed F₁). This analysis was performed after the completion of the testing stage where we fixed $\delta = 1.64$ (shown with red squares in Figure 2). Setting $\delta = 1.28$ would have enhanced the clustering output by 10% and therefore increased the final performance by 5%. The benefit of having a higher threshold is to be more certain that a given authorship link is correct, leading to higher clustering precisions. On the other hand, using a less restrictive threshold gives higher a clustering recall. We propose to be more cautious, mainly because proposing an incorrect assignment must be viewed as more problematic in many systems (especially if they are legal and law related) than missing a link between two documents written by the same author.

Interestingly, the authorship linking seems to produce a constant result (dashed line on the bottom, MAP) independent of the used threshold value δ.

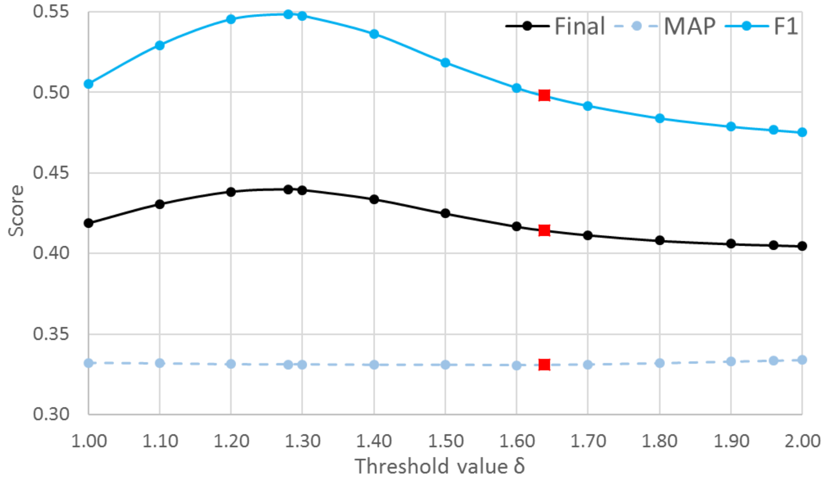


Figure 2. Results for various δ values after applying the bootstrap estimation.

Finally, we can evaluate the performance variation on the training data to determine the optimal length of the character *n*-grams for our second run. Figure 3 shows the mean clustering precision, recall, and the BCubed F_1 for different *n*-gram lengths from *n=1* (unigrams) to *n=12* based on the bootstrap approach. We can see a convergence from *n=2* to *n=9* between the recall (increasing) and the precision (decreasing) before they diverge again. In our second run, we used 6-grams (shown with red squares in Figure 3). The highest harmonic mean between precision and recall is achieved using 7-grams, which is only slightly better than the neighboring 6-grams and 8-grams (less than 0.5% change).

Overall, the analysis has shown that the chosen parameters are fine but could have been optimized. On the one hand, choosing Clark instead of Canberra as a distance measure or taking *n*-grams with length *n=7* characters instead of *n=6* characters would have unlikely improved the result noticeably. On the other hand, using a lower threshold value like $\delta = 1.28$ instead of $\delta = 1.64$ would have significantly enhanced the overall clustering performance.

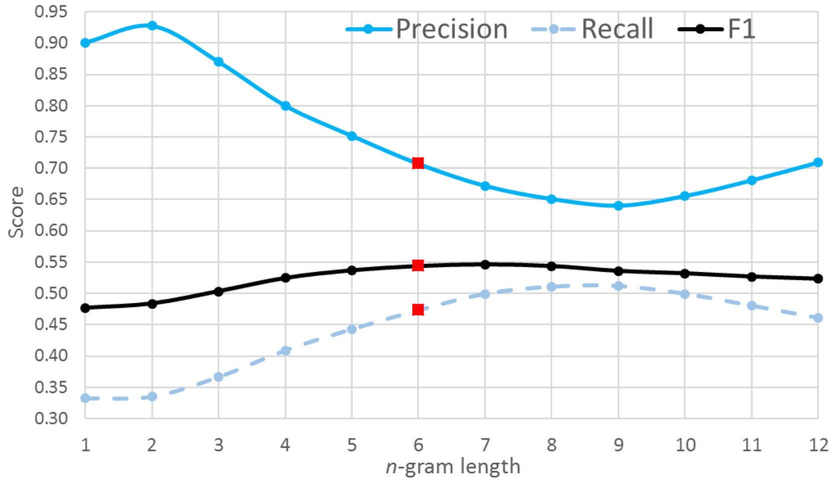


Figure 3. Results for various *n*-gram lengths after applying the bootstrap estimation.

6 Conclusion

This paper proposes a simple unsupervised technique to solve the author clustering problem. As features to discriminate between the proposed author and different candidates, we propose using the top *m* most frequent terms (words and punctuations) or character *n*-grams. This choice was found effective for other related tasks such as authorship attribution [2]. Moreover, compared to various feature selection strategies used in text categorization [12], the most frequent terms tend to select the most discriminative features when applied to stylistic studies [11]. To take the author linking decision, we propose using a simple distance measure called SPATIUM based on a variant of the L^1 norm called Canberra.

The proposed approach tends to perform very well in three different languages (Dutch, English, and Greek) and in two text genres (newspaper articles and reviews). Such a classifier strategy can be described as having a high bias but a low variance [4]. Changing the training data does not drastically change the decision. However, the suggested approach ignores other significant information such as mean sentence length, POS (part of speech) distribution, or topical terms. Even if the proposed system cannot capture all possible stylistic features (bias), changing the available data does not modify significantly the overall performance (variance).

It is common to fix some parameters (such as time period, size, genre, or length of the data) to minimize the possible sources of variation in the corpus. However, our main goal was to present a simple and unsupervised approach without too many predefined parameters.

With SPATIUM the proposed clustering decision could be clearly explained because it is based on a reduced set of features on the one hand and, on the other, those features are words, punctuation symbols, or long *n*-grams. Thus, the interpretation for the final user is clearer than when working with a huge number of features, when dealing with short *n*-grams of letters, or when combing several similarity measures. The SPATIUM decision can be explained by large differences in relative frequencies of frequent words, corresponding to either functional terms or overused topical words.

To improve the current classifier, we will investigate the consequence of other cluster linking strategies. Changing the single linkage strategy to a complete, average, or centroid linkage strategy could improve the outcome, because one sole link could no longer merge two bigger clusters and consequently not lower the precision drastically.

Acknowledgments. The author wants to thank the task coordinators for their valuable effort to promote test collections in author clustering. This research was supported, in part, by the NSF under Grant #200021_149665/1.

References

1. Amigo, E., Gonzalo, J., Artiles, J., & Verdejo, F. 2009. A comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints. *Information Retrieval*, 12(4), 461-486.
2. Burrows, J.F. 2002. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3), 267-287.
3. Gollub, T., Stein, B., & Burrows, T. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., & Sanderson, M. (eds.) SIGIR. *The 35th International ACM*, 1125–1126.
4. Hastie, T., Tibshirani, R., & Friedman, J. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer-Verlag: New York (NY).
5. Kocher, M., & Savoy, J. 2016. UniNE at CLEF 2016 Author Clustering: Notebook for PAN at CLEF 2016. In Balog, K., Capellato, L., Ferro, N., & Macdonald, C. (Eds), *CLEF 2016 Labs Working Notes, Évora, Portugal, September 5-8, 2016*, Aachen: CEUR.
6. Kocher, M., & Savoy, J. 2017. Author Clustering with an Adaptive Threshold. In Jones, G. J. F., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Thomas M., Cappellato, L., & Ferro, N. (Eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction, 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*. (to appear)
7. Kocher, M., & Savoy, J. 2017. Distance Measures in Author Profiling. *Information Processing & Management*, 53(5), 1103-1119.
8. Manning, C.D., Raghaven, P., & Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
9. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., & Stein, B. 2014. Improving the Reproducibility of PAN's Shared Tasks: - Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough,

P., Sanderson, M., Hall, M., Handbury, A., & Toms, E. (eds.) CLEF. *Lecture Notes in Computer Science*, vol. 8685, 268–299. Springer.

10. Savoy, J. 2016. Estimating the Probability of an Authorship Attribution. *Journal of American Society for Information Science & Technology*, 67(6), 1462-1472.
11. Savoy, J. 2015. Comparative Evaluation of Term Selection Functions for Authorship Attribution. *Digital Scholarship in the Humanities*, 30(2), 246-261.
12. Sebastiani, F. 2002. Machine Learning in Automatic Text Categorization. *ACM Computing Survey*, 34(1), 1-27.
13. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. 2016. Clustering by Authorship Within and Across Documents. In Working Notes of the CLEF 2016 Evaluation Labs, *CEUR Workshop Proceedings*, CEUR-WS.org.
14. Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M. 2017. Overview of the Author Identification Task at PAN 2017: Style Breach Detection and Author Clustering. In Working Notes Papers of the CLEF 2017 Evaluation Labs, *CEUR Workshop Proceedings*, CEUR-WS.org.