OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection

Notebook for PAN at CLEF 2017

Daniel Karaś, Martyna Śpiewak and Piotr Sobecki

National Information Processing Institute, Poland
{dkaras, mspiewak, psobecki}@opi.org.pl

Abstract. In this paper, we propose methods for author identification task dividing into author clustering and style breach detection. Our solution to the first problem consists of locality-sensitive hashing based clustering of real-valued vectors, which are mixtures of stylometric features and bag of n-grams. For the second problem, we propose a statistical approach based on some different tf-idf features that characterize documents. Applying the Wilcoxon Signed Rank test to these features, we determine the style breaches.

1 Author Clustering

1.1 Introduction

Author Clustering task consists of two distinct problems: author clustering and authorship link ranking. Solving first of the scenarios means assigning each of the m given documents to k clusters, where k is unknown and has to be approximated, where each of the k clusters corresponds to a single author. On the other hand, authorship link ranking can be understood as assigning intra-cluster confidence scores to document pairs, where a higher score indicates greater similarity between documents.

Both problems have to be solved for multiple collections of up to 50 documents. The additional difficulty lies in fact, that document batches were created in 3 different languages — English, Dutch, and Greek. This property makes it much harder to implement typical language-dependant solutions such as Word2Vec [3] or WodrNet [4], since such resources are not readily available for languages other than English. At its core, our solution to Author Clustering task consists of two main components: Locality-sensitive hashing (LSH) and Stylometric Measures that are not language-specific.

1.2 Locality-sensitive hashing

The goal of Local-sensitive hashing (LSH) is to cluster items into "buckets" by approximating similarities between aforementioned items. This group of algorithms is widely used in tasks such as clustering and near-duplicates detection.

There are multiple LSH algorithms. During our research we tested two of them — MinHash [9] and SuperBit [2]. After multiple evaluations, SuperBit proved to be better suited for described task. This algorithm approximates cosine similarity between

real-valued vectors and clusters them into given amount of clusters. The logic behind choosing this family of the algorithm is twofold: these algorithms have the reputation of being well suited for the task of clustering, we also wanted to test the tradeoff between their incredible speed and their effectiveness.

One of the main challenges of Author Clustering lies in establishing an optimal number of clusters since the count of clusters is not given a priori. Multiple solutions to this problem exist. Our final algorithm uses a process called silhouetting [11].

1.3 Stylometric Measures

Due to lack of language-dependant resources such as Word2Vec and WordNet for languages other than English, we decided to go with well known language-agnostic stylometric measures [5] as well as a typical bag of word n-grams representation. For the same reason — no stemming or lemmatization is performed on the documents.

Each document is represented as a fixed-size, real-valued vector. First part of the vector is a bag of word 3-grams, where each coordinate corresponds to unique word 3-gram present in a whole document collection for given problem.

For the rest of the vector, the mixture of multiple lexical word and character based measures are used. During the research, multiple different measures were evaluated, but at the end, we decided to use: special character frequency, average word length, average sentence length in characters, average sentence length in words and vocabulary richness (number of unique words divided by the number of words).

1.4 Results

Table 1: Results for PAN2017 training dataset

Problem	Language	Genre	F-Bcubed	R-Bcubed	"P-Bcubed	Av-Precision
problem001	en	articles	0.407890	0.344440	0.500000	0.032542
problem002	en	articles	0.383370	0.436670	0.341670	0.020267
problem003	en	articles	0.441710	0.354550	0.585710	0.031208
problem004	en	articles	0.494250	0.620000	0.410910	0.070715
problem005	en	articles	0.333330	1.000000	0.200000	0.127880
problem006	en	articles	0.600000	0.866670	0.458820	0.277360
problem007	en	articles	0.393570	1.000000	0.245000	0.235450
problem008	en	articles	0.731530	0.661110	0.818750	0.485970
problem009	en	articles	0.389530	0.363890	0.419050	0.023356
problem010	en	articles	0.428910	0.319050	0.654170	0.105910
problem011	en	reviews	0.473870	0.421150	0.541670	0.114890
problem012	en	reviews	0.677000	0.753330	0.614710	0.346780
problem013	en	reviews	0.473630	0.853330	0.327780	0.170070
problem014	en	reviews	0.405570	0.366670	0.453700	0.043251
problem015	en	reviews	0.509020	0.658930	0.414680	0.168070

Continued on next page

Table 1 – *Results for PAN2017 training dataset*

Problem	Language	Genre	F-Bcubed	R-Bcubed	"P-Bcubed	Av-Precision
problem016	en	reviews	0.405020	0.600480	0.305560	0.142210
problem017	en	reviews	0.408400	0.443330	0.378570	0.065487
problem018	en	reviews	0.554640	0.493330	0.633330	0.028054
problem019	en	reviews	0.375870	0.789290	0.246670	0.179890
problem020	en	reviews	0.353110	0.820000	0.225000	0.070972
problem021	nl	articles	0.495550	0.497780	0.493330	0.063403
problem022	nl	articles	0.461920	0.387140	0.572500	0.094984
problem023	nl	articles	0.400250	0.735000	0.275000	0.073250
problem024	nl	articles	0.515130	0.518180	0.512120	0.219490
problem025	nl	articles	0.524570	0.733330	0.408330	0.125440
problem026	nl	articles	0.559890	0.446670	0.750000	0.170080
problem027	nl	articles	0.360600	0.457140	0.297730	0.042885
problem028	nl	articles	0.429240	0.420000	0.438890	0.032622
problem029	nl	articles	0.598770	0.746150	0.500000	0.273150
problem030	nl	articles	0.504400	0.426190	0.617780	0.147790
problem031	nl	reviews	0.497900	0.781250	0.365380	0.252900
problem032	nl	reviews	0.523900	0.468750	0.593750	0.078873
problem033	nl	reviews	0.412700	0.361110	0.481480	0.002976
problem034	nl	reviews	0.515000	0.678570	0.414970	0.178020
problem035	nl	reviews	0.474580	0.400000	0.583330	0.132480
problem036	nl	reviews	0.469260	0.416670	0.537040	0.004902
problem037	nl	reviews	0.322500	0.600000	0.220510	0.151300
problem038	nl	reviews	0.535290	0.433330	0.700000	0.028499
problem039	nl	reviews	0.463160	0.400000	0.550000	0.000000
problem040	nl	reviews	0.432780	0.683330	0.316670	0.196850
problem041	gr	articles	0.425240	0.636670	0.319230	0.090813
problem042	gr	articles	0.478660	0.595830	0.400000	0.131320
problem043	gr	articles	0.520610	0.761670	0.395450	0.163680
problem044	gr	articles	0.493880	0.728330	0.373610	0.197920
problem045	gr	articles	0.415200	0.520000	0.345560	0.042738
problem046	gr	articles	0.519860	0.700000	0.413460	0.171660
problem047	gr	articles	0.453640	0.691670	0.337500	0.163980
problem048	gr	articles	0.479610	0.660000	0.376670	0.102200
problem049	gr	articles	0.470300	0.500000	0.443940	0.108860
problem050	gr	articles	0.449520	0.383330	0.543330	0.131710
problem051	gr	reviews	0.480540	0.420830	0.560000	0.055130
problem052	gr	reviews	0.393060	0.636670	0.284290	0.093994
problem053	gr	reviews	0.534860	0.567500	0.505770	0.182710
problem054	gr	reviews	0.459390	0.551110	0.393850	0.105250
problem055	gr	reviews	0.509330	0.916670	0.352630	0.237980
problem056	gr	reviews	0.394480	0.593330	0.295450	0.042487

Continued on next page

Table 1 – *Results for PAN2017 training dataset*

Problem	Language	Genre	F-Bcubed	R-Bcubed	"P-Bcubed	Av-Precision
problem057	gr	reviews	0.365170	0.596670	0.263100	0.038210
problem058	gr	reviews	0.461150	0.437500	0.487500	0.063835
problem059	gr	reviews	0.515050	0.745830	0.393330	0.109060
problem060	gr	reviews	0.483030	0.630000	0.391670	0.044900

Table 2: Results for PAN2017 test dataset

Problem	Language	Genre	F-Bcubed	R-Bcubed	"P-Bcubed	Av-Precision
problem001	en	articles	0.645930	0.696670	0.602080	0.400580
problem002	en	articles	0.463950	0.383330	0.587500	0.081134
problem003	en	articles	0.418680	0.461900	0.382860	0.124740
problem004	en	articles	0.412690	0.543330	0.332690	0.083299
problem005	en	articles	0.628290	0.623330	0.633330	0.282090
problem006	en	articles	0.418510	0.398330	0.440830	0.060129
problem007	en	articles	0.423770	0.348720	0.540000	0.072016
problem008	en	articles	0.482420	0.460000	0.507140	0.079461
problem009	en	articles	0.776280	0.738890	0.817650	0.474400
problem010	en	articles	0.572720	0.516670	0.642420	0.165370
problem011	en	articles	0.462030	0.424290	0.507140	0.014544
problem012	en	articles	0.528660	0.575000	0.489230	0.123790
problem013	en	articles	0.450820	0.644440	0.346670	0.092703
problem014	en	articles	0.621250	0.633330	0.609620	0.205350
problem015	en	articles	0.424140	0.552380	0.344230	0.027974
problem016	en	articles	0.479660	0.658330	0.377270	0.154390
problem017	en	articles	0.487220	0.458330	0.520000	0.029075
problem018	en	articles	0.520000	0.433330	0.650000	0.022727
problem019	en	articles	0.446230	0.543330	0.378570	0.072511
problem020	en	articles	0.490040	0.485710	0.494440	0.100070
problem021	en	reviews	0.345450	0.950000	0.211110	0.221300
problem022	en	reviews	0.350800	0.512500	0.266670	0.066592
problem023	en	reviews	0.353910	1.000000	0.215000	0.272140
problem024	en	reviews	0.400190	0.600830	0.300000	0.116170
problem025	en	reviews	0.337180	0.875000	0.208820	0.065738
problem026	en	reviews	0.469780	0.508330	0.436670	0.034419
problem027	en	reviews	0.402840	0.522220	0.327880	0.053262
problem028	en	reviews	0.494430	0.600000	0.420450	0.040009
problem029	en	reviews	0.501390	0.720000	0.384620	0.061785
problem030	en	reviews	0.380680	0.860000	0.244440	0.078936
problem031	en	reviews	0.321360	0.218570	0.606670	0.021624
problem032	en	reviews	0.492580	0.673330	0.388330	0.227330
problem033	en	reviews	0.516360	0.708330	0.406250	0.153760

Continued on next page

Table 2 – *Results for PAN2017 test dataset*

Problem	Language	Genre	F-Bcubed	R-Bcubed	"P-Bcubed	Av-Precision
problem034	en	reviews	0.330710	0.230770	0.583330	0.011269
problem035	en	reviews	0.567830	0.578330	0.557690	0.187770
problem036	en	reviews	0.487760	0.697500	0.375000	0.114990
problem037	en	reviews	0.384690	0.415240	0.358330	0.033317
problem038	en	reviews	0.431790	0.369440	0.519440	0.051175
problem039	en	reviews	0.512680	0.407690	0.690480	0.070637
problem040	en	reviews	0.470650	0.834290	0.327780	0.218080
problem041	nl	articles	0.368420	0.233330	0.875000	0.136820
problem042	nl	articles	0.406060	0.574170	0.314100	0.190250
problem043	nl	articles	0.486960	0.700000	0.373330	0.270560
problem044	nl	articles	0.398870	0.371110	0.431110	0.063250
problem045	nl	articles	0.636000	0.750000	0.552080	0.313590
problem046	nl	articles	0.465100	0.390480	0.575000	0.074737
problem047	nl	articles	0.387770	0.320830	0.490000	0.020058
problem048	nl	articles	0.461540	1.000000	0.300000	0.369360
problem049	nl	articles	0.498750	0.525000	0.475000	0.105010
problem050	nl	articles	0.468940	0.342860	0.741670	0.059018
problem051	nl	articles	0.405010	0.397780	0.412500	0.102280
problem052	nl	articles	0.427850	0.466670	0.395000	0.018594
problem053	nl	articles	0.548000	0.694440	0.452560	0.228040
problem054	nl	articles	0.517640	0.637500	0.435710	0.116780
problem055	nl	articles	0.439160	0.563890	0.359620	0.046793
problem056	nl	articles	0.421090	0.440480	0.403330	0.062252
problem057	nl	articles	0.561150	1.000000	0.390000	0.332110
problem058	nl	articles	0.473750	0.620000	0.383330	0.162460
problem059	nl	articles	0.486730	0.533330	0.447620	0.042951
problem060	nl	articles	0.368980	0.415000	0.332140	0.050216
problem061	nl	reviews	0.498180	0.875000	0.348210	0.241950
problem062	nl	reviews	0.444350	0.712960	0.322750	0.134440
problem063	nl	reviews	0.377590	0.500000	0.303330	0.167750
problem064	nl	reviews	0.411470	0.468750	0.366670	0.042372
problem065	nl	reviews	0.443180	0.375000	0.541670	0.015341
problem066	nl	reviews	0.418950	0.593750	0.323660	0.157740
problem067	nl	reviews	0.533770	0.453700	0.648150	0.084187
problem068	nl	reviews	0.402930	0.333330	0.509260	0.017273
problem069	nl	reviews	0.466600	0.472220	0.461110	0.074354
problem070	nl	reviews	0.506730	0.531250	0.484380	0.073393
problem071	nl	reviews	0.517940	0.540000	0.497620	0.042484
problem072	nl	reviews	0.549850	0.583330	0.520000	0.095238
problem073	nl	reviews	0.545450	0.500000	0.600000	0.071429
problem074	nl	reviews	0.444440	0.400000	0.500000	0.000000

Continued on next page

Table 2 – *Results for PAN2017 test dataset*

Problem	Language	Genre	F-Bcubed	R-Bcubed	"P-Bcubed	Av-Precision
problem075	nl	reviews	0.515460	0.550000	0.485000	0.057313
problem076	nl	reviews	0.497810	0.775000	0.366670	0.109060
problem077	nl	reviews	0.574230	0.650000	0.514290	0.054895
problem078	nl	reviews	0.445810	0.475000	0.420000	0.004762
problem079	nl	reviews	0.294550	0.858330	0.177780	0.122440
problem080	nl	reviews	0.582650	0.483330	0.733330	0.010417
problem081	gr	articles	0.500940	0.900000	0.347060	0.183760
problem082	gr	articles	0.479490	0.425000	0.550000	0.051942
problem083	gr	articles	0.562320	0.584620	0.541670	0.235550
problem084	gr	articles	0.454610	0.541670	0.391670	0.080786
problem085	gr	articles	0.404820	0.491670	0.344050	0.099311
problem086	gr	articles	0.365330	0.454170	0.305560	0.123180
problem087	gr	articles	0.317580	0.504760	0.231670	0.035122
problem088	gr	articles	0.523250	0.710000	0.414290	0.066747
problem089	gr	articles	0.795180	1.000000	0.660000	0.536570
problem090	gr	articles	0.662110	0.825000	0.552940	0.338080
problem091	gr	articles	0.650880	0.620000	0.685000	0.352410
problem092	gr	articles	0.519040	0.857140	0.372220	0.277550
problem093	gr	articles	0.544930	0.526320	0.564910	0.111680
problem094	gr	articles	0.496540	0.610710	0.418330	0.198450
problem095	gr	articles	0.383130	0.530000	0.300000	0.131270
problem096	gr	articles	0.407150	0.291730	0.673680	0.035518
problem097	gr	articles	0.577060	0.444120	0.823610	0.285260
problem098	gr	articles	0.429490	0.775000	0.297060	0.165900
problem099	gr	articles	0.457100	0.737140	0.331250	0.166660
problem100	gr	articles	0.435760	0.441670	0.430000	0.039691
problem101	gr	reviews	0.365790	0.566900	0.270000	0.143080
problem102	gr	reviews	0.405040	0.434440	0.379370	0.070181
problem103	gr	reviews	0.419470	0.733330	0.293750	0.132500
problem104	gr	reviews	0.495240	0.650000	0.400000	0.154300
problem105	gr	reviews	0.515040	0.557140	0.478850	0.123990
problem106	gr	reviews	0.495700	0.708330	0.381250	0.099837
problem107	gr	reviews	0.485800	0.440380	0.541670	0.145940
problem108	gr	reviews	0.426770	0.640480	0.320000	0.213520
problem109	gr	reviews	0.452050	0.361430	0.603330	0.200960
problem110	gr	reviews	0.377070	0.583330	0.278570	0.155800
problem111	gr	reviews	0.384740	0.775000	0.255880	0.093850
problem112	gr	reviews	0.430200	0.500000	0.377500	0.066225
problem113	gr	reviews	0.397240	0.622620	0.291670	0.181710
problem114	gr	reviews	0.356770	0.716670	0.237500	0.058730
problem115	gr	reviews	0.374060	0.808330	0.243330	0.097110

Continued on next page

Table 2 – *Results for PAN2017 test dataset*

Problem	Language	Genre	F-Bcubed	R-Bcubed	"P-Bcubed	Av-Precision
problem116	gr	reviews	0.430040	0.640000	0.323810	0.119100
problem117	gr	reviews	0.431020	0.478330	0.392220	0.008253
problem118	gr	reviews	0.452780	0.737500	0.326670	0.108080
problem119	gr	reviews	0.420360	0.795000	0.285710	0.085212
problem120	gr	reviews	0.470340	0.620830	0.378570	0.145050

1.5 Method summary

Our solution to author clustering and authorship link can be written in following steps: First, we approximate the desired amount of clusters using silhouetting, then we represent every document in a collection as a real-valued vector consisting of a bag of word 3-grams and multiple stylometric measures, then SuperBit LSH algorithms is used for the actual clustering procedure. Authorship link is calculated using cosine similarity.

2 Style Breach Detection

2.1 Introduction

Style Breach Detection task consists in detecting borders where authorship may change within a document. Unlike the text segmentation problem which mainly focuses on finding switches of topics, whereas the point of style breach detection task lies in discovering borders using writing style features ignoring analysis the content of the text.

We propose a statistical approach based on tf-idf features that characterize documents from widely different points of view: word n-grams (we consider only $n = 1$ and $n = 3$), punctuation, Part of Speech (PoS) using The Penn Treebank POS Tagger [12], stopwords, to determine the borders of changing style within a document.

2.2 The Wilcoxon Signed Rank Test

The paired samples Wilcoxon signed-rank test is a nonparametric test which is used to verify the null hypothesis that two samples come from the same distribution [1].

Suppose we have a random sample of N pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$, where $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ correspond to the blocks/objects effect before and after some activity, respectively. For each random sample the difference is formed as $D_i = X_i - Y_i$. We assume the observation $D_1, \ldots, D_N$ are independent from a population which is continuous and symmetric with median M_D. We verify the null hypothesis $H_0 : M_D = 0$ against the two-sided alternative $H_1 : M_D \neq 0$.

The algorithm to determine the statistic of this test is as follows: we need to order the absolute differences $|D_1|, \ldots, |D_n|$ from the smallest to the largest and assign them N integer ranks (from 1 to N), noting the original signs of the differences D_i. We consider the sum of ranks of the positive differences as a test criterion because the sum

of all the ranks is a constant. If we denote r as the rank of a random variable, then the test statistic can be written as

$$T = \sum_{i=1}^{n} r(|D_i|) I(D_i > 0), \tag{1}$$

where $I(\rho) = 1$ if a sentence ρ is true and $I(\rho) = 0$ otherwise.

We denote Z_i by $I(D_i > 0)$ for each $i = 1, \ldots, N$. Under the null hypothesis the Z_i are independent and identically distributed from Bernoulli population with probability $P(Z_i = 1) = \frac{1}{2}$. The test statistic is a linear combination of Z_i variables, so we could determine its expected value and variance as follows:

$$\mathrm{E}(T) = \frac{n(n+1)}{4}, \tag{2}$$

$$\mathrm{Var}(T) = \frac{n(n+1)(2n+1)}{24}. \tag{3}$$

We apply approximation based on the asymptotic normality of T due to lack of knowledge the exact distribution of this statistic. The following statistic:

$$T^* = \frac{T - \mathrm{E}(T)}{\sqrt{\mathrm{Var}(T)}} \tag{4}$$

is asymptotically normal under H_0.

Let α denote an accepted significance level. We reject the null hypothesis against the two-sided alternative if $|T^*| \geq z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)^{\text{th}}$ quantile from a normal distribution with mean 0 and standard deviation 1.

2.3 Tf-idf: Term frequency–inverse document frequency

Originally, tf-idf calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in [10].

Formally, tf-idf is the product of term frequency and inverse document frequency. The term frequency is the number of times that i-th word occurs in j-th document, and it may be written as

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{5}$$

where $n_{i,j}$ is the number of occurrences the i-th word in the j-th document and the denominator is the sum of the number of occurrences of all words in the j-th document. The inverse document frequency is the logarithm of the inverse fraction of the documents that contain the i-th word:

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d : w_i \in d\}|}, \tag{6}$$

where $|D|$ is the number of all documents in the given corpus and the denominator is equal to the number of documents where the i-th word occurs at least once. Then, tf-idf for i-th word and the j-th document is as follows:

$$\text{tf-idf}_{i,j} = \mathrm{tf}_{i,j} \cdot \mathrm{idf}_i. \tag{7}$$

2.4 The paired samples Wilcoxon Signed Rank test with tf-idf features to detect style breaches

The corpus used to construct our approach consists of only documents that are provided in English and may contain either zero or many style breaches which occur at the end sentences. Further, we noticed paragraphs are natural borders of the style breaches. On this account, we split each document into sections assuming nothing less than two blank lines determine the boundary between two paragraphs. If there are not any blank lines within a document, then m sentences are organized into a section, where m is a fixed natural number.

Customarily, tf–idf is a numerical statistic that is intended to reflect how important a word is to a document in a corpus [9]. In our approach, we use tf-idf to determine how important a particular term is to a paragraph in a document. For each document and each term mentioned above, we determine the tf-idf matrix X_i, where we denote X_1, X_2, X_3, X_4, X_5 as the tf-idf matrix for word, punctuation, PoS, stopwords, word 3-grams, respectively. The number of rows of X_i is equal to the number of paragraphs in a document, and the number of columns of this matrix is equal to the number of all unique terms in this document.

We computed vectors representing paragraphs as concatenated tf-idf vectors of selected terms together, it may be written as:

$$x_k = (x_{k,j_1}, \dots x_{k,j_s}), \qquad (j_1, \dots, j_s) \subset \{1, , 5\}, \tag{8}$$

where we denote x_k as tf-idf combining vector for the k-th paragraph as concatenated s tf-idf vectors of above-mentioned terms together ($x_{k,j}$ is tf-idf vector of the j-th term for the k-th paragraph).

The primary aim of this approach is to test whether one or multi-authors wrote two following paragraphs. For this purpose, we use the paired samples Wilcoxon Signed Rank test which is used to verify if two samples come from the same distribution. We assume if the same author write two paragraphs they should have the same distribution and analogously if two paragraphs are not written by the same author they come from the different distributions. In other words, if the same author has drafted two sections the result of the test should not be statistically significant (the null hypothesis is accepted, the style is not changing between two consecutive paragraphs). On the other hand, if multi-authors write two paragraphs then the null hypothesis should be rejected (the style difference between two sections is statistically significant).

For each two consecutive paragraphs in a document, we test if these paragraphs have the same style. As the result of these tests, we note p-values. Next, we sort the p-values from smallest to largest value, and we determine the S lowest p-values, where S is defined as:

$$S = \lfloor p \cdot |P| \rfloor + 1, \tag{9}$$

where p is a fixed value that lies in $[0, 1]$ and $|P|$ is the number of paragraphs in a document.

The borders between paragraphs corresponding with selected p-values imply the style breaches.

Table 3. Results for training evaluations according to subsets of tf-idf features, m and p are fixed ($m = 10$ and $p = 0.3$).

Combine features	**WindowDiff**	**WinP**	**WinR**	**WinF**
$[X_2, X_4, X_5]$	0.526434	0.344312	0.620210	0.342847
$[X_4, X_5]$	0.527448	0.343365	0.619061	0.341818
$[X_3, X_4, X_5]$	0.525729	0.343161	0.617870	0.341496
$[X_2, X_3, X_4, X_5]$	0.526380	0.341384	0.616980	0.340310
$[X_1, X_4]$	0.534279	0.339005	0.617799	0.337278
[X_1, X_3, X_4]	**0.535459**	**0.336084**	**0.612633**	**0.333563**
$[X_1, X_3, X_5]$	0.532014	0.332278	0.613327	0.333267
$[X_1, X_5]$	0.532141	0.331560	0.614199	0.333213
$[X_5]$	0.533403	0.333675	0.610578	0.333127
$[X_1, X_2, X_3, X_5]$	0.532065	0.332111	0.613266	0.333075
$[X_1, X_2, X_4, X_5]$	0.534239	0.331392	0.613619	0.332709
$[X_1, X_4, X_5]$	0.533170	0.331391	0.613619	0.332707
$[X_1, X_2, X_5]$	0.532724	0.331013	0.613450	0.332558
$[X_2, X_3, X_5]$	0.534358	0.332229	0.610030	0.332168
$[X_3, X_5]$	0.533902	0.332250	0.609848	0.332164
$[X_2, X_5]$	0.534230	0.331857	0.609818	0.331915
$[X_1, X_2, X_3, X_4]$	0.536221	0.334333	0.610863	0.331759
$[X_1, X_3]$	0.534803	0.326948	0.615239	0.331622
$[X_2, X_4]$	0.537484	0.332465	0.615676	0.331344
$[X_1, X_3, X_4, X_5]$	0.534146	0.330113	0.611450	0.331102
$[X_1, X_2, X_3, X_4, X_5]$	0.534129	0.330113	0.611450	0.331102
$[X_1, X_2, X_4]$	0.538384	0.332572	0.611185	0.330859
$[X_3]$	0.539429	0.327774	0.608097	0.330372
$[X_4]$	0.534342	0.331434	0.609035	0.329315
$[X_1, X_2]$	0.541324	0.325407	0.611997	0.328848
$[X_1, X_2, X_3]$	0.537711	0.323988	0.612137	0.328599
$[X_1]$	0.541647	0.321542	0.609613	0.325763
$[X_2, X_3]$	0.540967	0.322548	0.606527	0.325516
$[X_2, X_3, X_4]$	0.542909	0.326722	0.605585	0.324760
$[X_3, X_4]$	0.541420	0.326358	0.603533	0.323806
$[X_2]$	0.561822	0.312071	0.599578	0.315449

2.5 Evaluations and Results

The main goal of training evaluations was to choose the set of values of the parameters used in our submitted solution. Keeping in mind the previous PAN's task — Intrinsic Plagiarism Detection task [8], we assumed that at least of 70% of each document was written by the one primary author, other 30% of a text could be written by other authors, eventually. Hence we fixed p as 0.3. Additionally, our initial experiments showed that best results were obtained for $m = 10$.

Therefore, the principal evaluation to determine the optimal set of tf-idf features we performed for the parameters mentioned above. In Table 3, we showed the detailed results according to the subset of tf-idf features. It worth noticing that our primary

Table 4. Official results for PAN2017 test dataset

Team	winF	winP	winR	windowDiff	Runtime
OPI-JSA	**0.322601**	0.314656	**0.585617**	0.545648	**00:01:19**
khan17	0.288795	**0.399004**	0.487075	**0.479990**	00:02:23
kuznetsova17	0.277264	0.371108	0.542527	0.529496	00:20:25

intention was optimized the F-score of WinPR. Due to the similar results obtained on the training dataset, we select the subset of tf-idf features which also gives good results on other datasets, based on our previous experiences. For the final submission, we chose tf-idf of word, PoS and stopwords.

In Table 4, the official results were shown [6]. Our submitted solution took the first place according to winF, winR, and runtime. The proposed approach optimizes recall at the sacrifice of precision and windowDiff (what was the main intention of our system).

3 Conclusion

We have presented methods for author identification task [13] that we submitted to the 2017 PAN competition [7]. This year the author identification task was divided into author clustering and style breach detection tasks. We proposed solutions for these competitions independently.

The submitted system for style breach detection task obtained the best result according to F-score of WinPR that it uses for the final ranking of all participating teams. Additionally, it is worth noticing we were building both of our algorithms bearing in mind optimizing execution time. Both systems had the shortest runtimes of all submitted solutions. Implementation of our solution of author clustering task achieved the fastest running time, which could be further improved if the number of clusters would be known a priori for each problem, since the routine of optimizing number of clusters for each problem is the most time-consuming step of the algorithm. While exhibiting remarkable running time, our algorithm did not perform substantially worse than other contestants. For the kind of usage cases that we are going to employ said algorithm for — the trade-off between running time and performance proved to be satisfying, which means we may use it in real-world scenarios after few improvements like using language-specific tools such as WordNet.

References

1. Gibbons, J.D., Chakraborti, S.: Nonparametric statistical inference. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science, pp. 977–979. Springer (2011)
2. Ji, J., Li, J., Yan, S., Zhang, B., Tian, Q.: Super-bit locality-sensitive hashing. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 108–116. Curran Associates, Inc. (2012), http://papers.nips.cc/paper/4847-super-bit-locality-sensitive-hashing.pdf
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), http://arxiv.org/abs/1301.3781

4. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (Nov 1995), http://doi.acm.org/10.1145/219717.219748
5. Pervaz, I., Ameer, I., Sittar, A., Nawab, R.M.A.: Identification of author personality traits using stylistic features: Notebook for pan at clef 2015. In: Cappellato, L., Ferro, N., Jones, G.J.F., SanJuan, E. (eds.) CLEF (Working Notes). CEUR Workshop Proceedings, vol. 1391. CEUR-WS.org (2015), http://dblp.uni-trier.de/db/conf/clef/clef2015w.htmlPervazASN15
6. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: 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., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
7. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17). Springer, Berlin Heidelberg New York (Sep 2017)
8. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 1–9. CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
9. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press (2011)
10. Ramos, J.: Using tf-idf to determine word relevance in document queries (2003)
11. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53 – 65 (1987), http://www.sciencedirect.com/science/article/pii/0377042787901257
12. Santorini, B.: Part-of-speech tagging guidelines for the Penn Treebank Project. Tech. Rep. MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania (1990), ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz
13. Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs