Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features

PAN

CLEF 2019 Conference and Labs of the Evaluation Forum - Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 9 - 12 September 2019, Lugano

Andrea Bacciu, Massimo La Morgia, , Alessandro Mei, Eugenio N. Nemmi, Valerio Neri, Julinda Stefa.

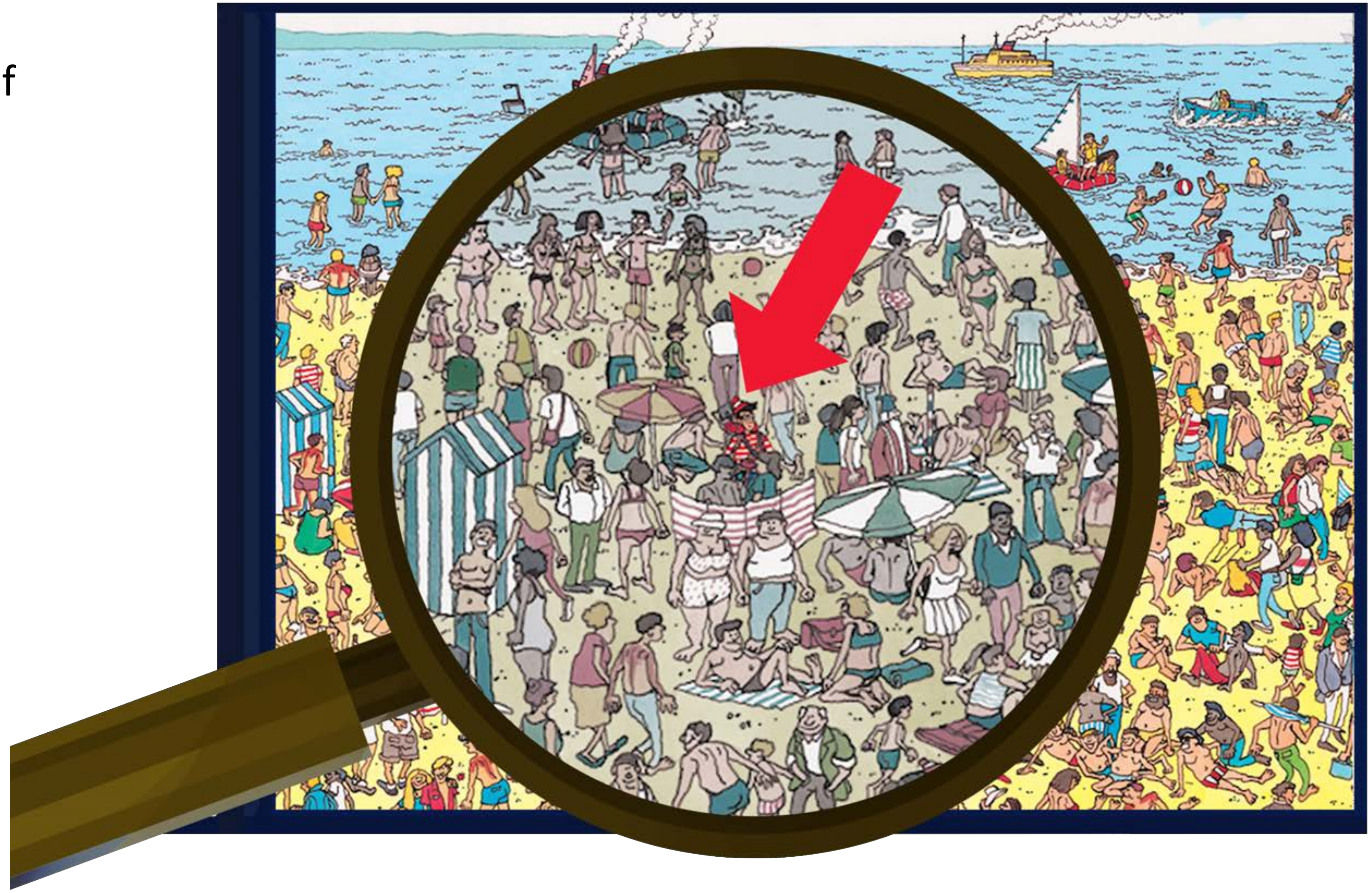
Speaker

Eugenio N. Nemmi

SAPIENZA
UNIVERSITÀ DI ROMA

PAN 2019 Authorship Attribution Task

- **Authorship attribution** is the task of identifying the **author** of a given text.

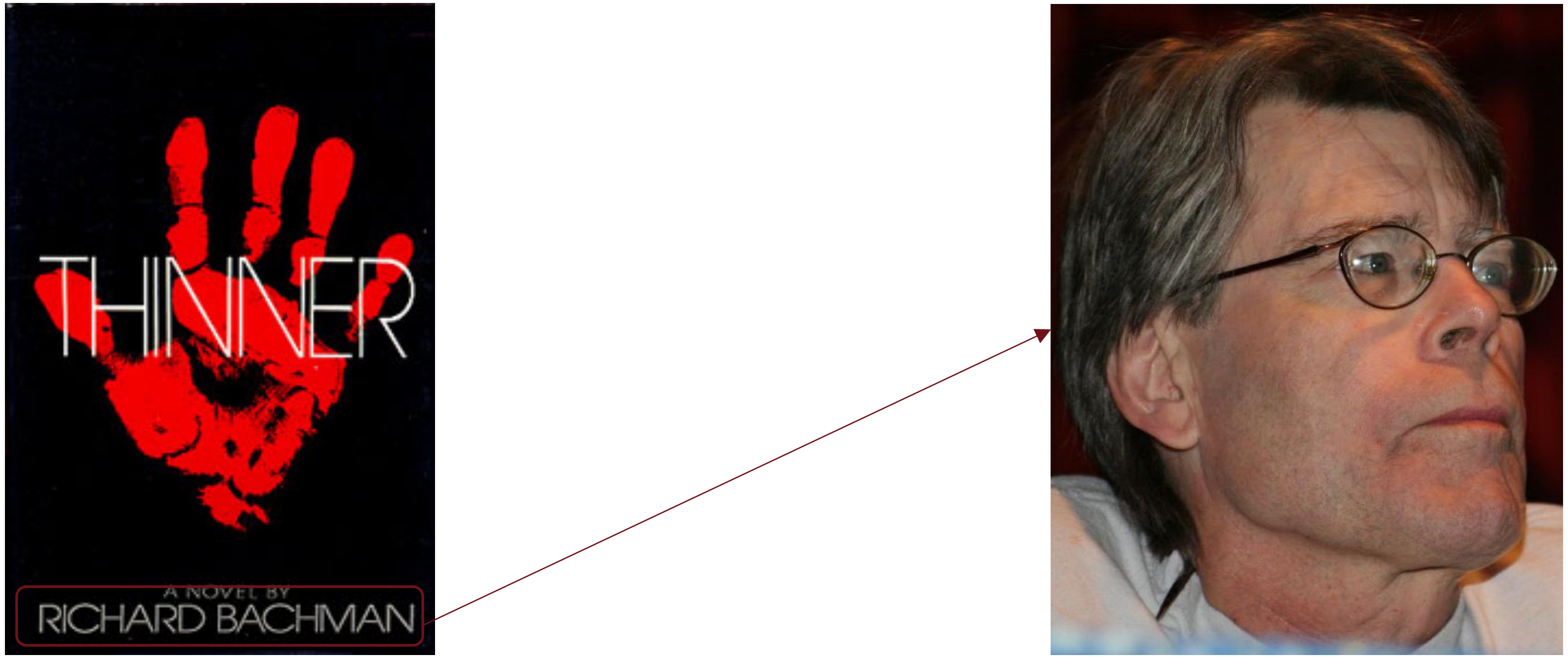


Sapienza
Università di Roma

Motivation

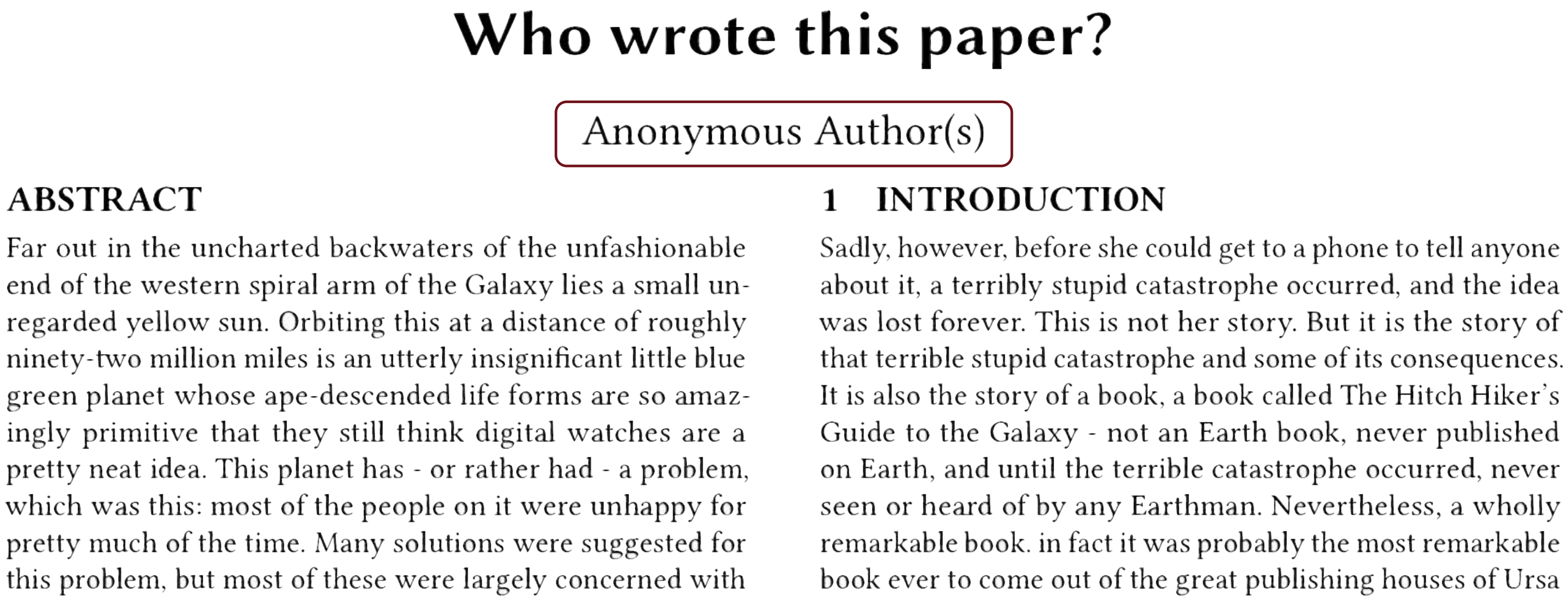
Detect real author of a novel

Stephen King



Motivation

Detect author of paper in double blind review submission



Who wrote this paper?

Anonymous Author(s)

ABSTRACT

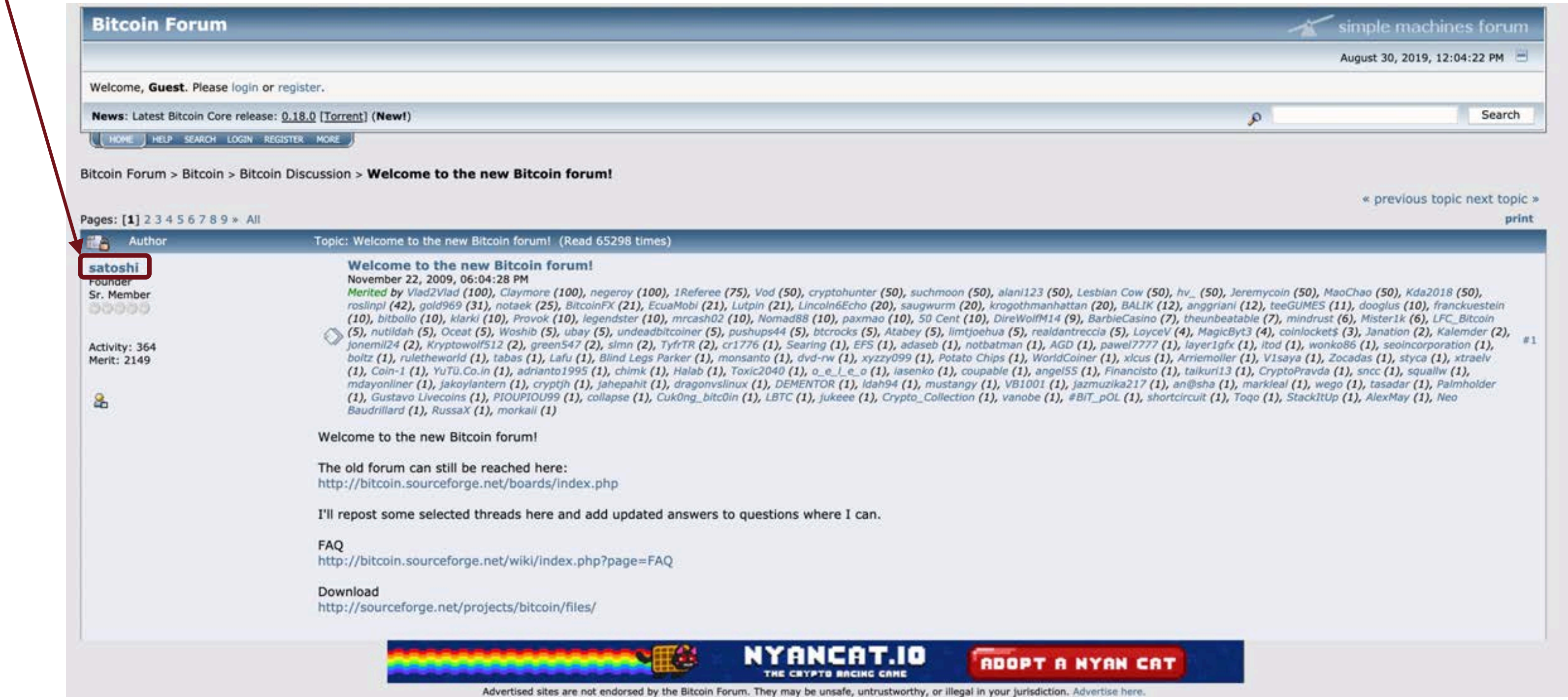
Far out in the uncharted backwaters of the unfashionable end of the western spiral arm of the Galaxy lies a small unregarded yellow sun. Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue green planet whose ape-descended life forms are so amazingly primitive that they still think digital watches are a pretty neat idea. This planet has - or rather had - a problem, which was this: most of the people on it were unhappy for pretty much of the time. Many solutions were suggested for this problem, but most of these were largely concerned with

1 INTRODUCTION

Sadly, however, before she could get to a phone to tell anyone about it, a terribly stupid catastrophe occurred, and the idea was lost forever. This is not her story. But it is the story of that terrible stupid catastrophe and some of its consequences. It is also the story of a book, a book called The Hitch Hiker's Guide to the Galaxy - not an Earth book, never published on Earth, and until the terrible catastrophe occurred, never seen or heard of by any Earthman. Nevertheless, a wholly remarkable book. in fact it was probably the most remarkable book ever to come out of the great publishing houses of Ursa

Motivation

Deanonymize Pseudonyms



Bitcoin Forum

August 30, 2019, 12:04:22 PM

Welcome, **Guest**. Please login or register.

News: Latest Bitcoin Core release: 0.18.0 [Torrent] **(New!)**

Bitcoin Forum > Bitcoin > Bitcoin Discussion > **Welcome to the new Bitcoin forum!**

« previous topic next topic »

Pages: [**1**] 2 3 4 5 6 7 8 9 » All

print

Author	Topic: Welcome to the new Bitcoin forum! (Read 65298 times)
satoshi Founder Sr. Member Activity: 364 Merit: 2149	**Welcome to the new Bitcoin forum!** November 22, 2009, 06:04:28 PM *Merited by Vlad2Vlad (100), Claymore (100), negeroy (100), 1Referee (75), Vod (50), cryptohunter (50), suchmoon (50), alani123 (50), Lesbian Cow (50), hv_ (50), Jeremycoin (50), MaoChao (50), Kda2018 (50), roslinpl (42), gold969 (31), notaek (25), BitcoinFX (21), EcuaMobi (21), Lutpin (21), Lincoln6Echo (20), saugwurm (20), krogothmanhattan (20), BALIK (12), anggriani (12), teeGUMES (11), dooglus (10), franckuestein (10), bitbollo (10), klarki (10), Provok (10), legendster (10), mrcash02 (10), Nomad88 (10), paxmao (10), 50 Cent (10), DireWolfM14 (9), BarbieCasino (7), theunbeatable (7), mindrust (6), Mister1k (6), LFC_Bitcoin (5), nutildah (5), Oceat (5), Woshib (5), ubay (5), undeadbitcoiner (5), pushups44 (5), btcrocks (5), Atabey (5), limtjoehua (5), realdantreccia (5), LoyceV (4), MagicByt3 (4), coinlocket$ (3), Janation (2), Kalemder (2), jonemil24 (2), Kryptowolf512 (2), green547 (2), slmn (2), TyfrTR (2), cr1776 (1), Searing (1), EFS (1), adaseb (1), notbatman (1), AGD (1), pawel7777 (1), layer1gfx (1), itod (1), wonko86 (1), seoincorporation (1), boltz (1), ruletheworld (1), tabas (1), Lafu (1), Blind Legs Parker (1), monsanto (1), dvd-rw (1), xyzzy099 (1), Potato Chips (1), WorldCoiner (1), xlcus (1), Arriemoller (1), V1saya (1), Zocadas (1), styca (1), xtraelv (1), Coin-1 (1), YuTü.Co.in (1), adrianto1995 (1), chimk (1), Halab (1), Toxic2040 (1), o_e_l_e_o (1), iasenko (1), coupable (1), angel55 (1), Financisto (1), taikuri13 (1), CryptoPravda (1), sncc (1), squallw (1), mdayonliner (1), jakoylantern (1), cryptjh (1), jahepahit (1), dragonvslinux (1), DEMENTOR (1), ldah94 (1), mustangy (1), VB1001 (1), jazmuzika217 (1), an@sha (1), markleal (1), wego (1), tasadar (1), Palmholder (1), Gustavo Livecoins (1), PIOUPIOU99 (1), collapse (1), Cuk0ng_bitc0in (1), LBTC (1), jukeee (1), Crypto_Collection (1), vanobe (1), #BiT_pOL (1), shortcircuit (1), Toqo (1), StackItUp (1), AlexMay (1), Neo Baudrillard (1), RussaX (1), morkaii (1)* Welcome to the new Bitcoin forum! The old forum can still be reached here: http://bitcoin.sourceforge.net/boards/index.php I'll repost some selected threads here and add updated answers to questions where I can. FAQ http://bitcoin.sourceforge.net/wiki/index.php?page=FAQ Download http://sourceforge.net/projects/bitcoin/files/

Unknown Text

Fifotofotofoto donono dodod did idodid odofofof ififododi.

Wooooooa!

Noot noot!

AA Scenarios

Closed-set

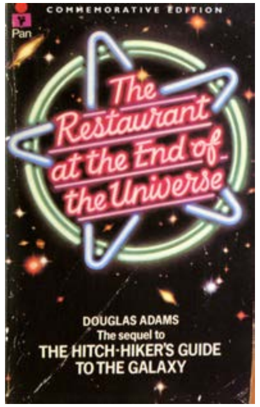
Finite set of candidates authors among which there is the real author.

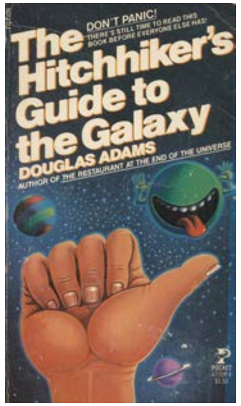
Open-set

The author of a disputed text is not necessarily included in the list of candidates.

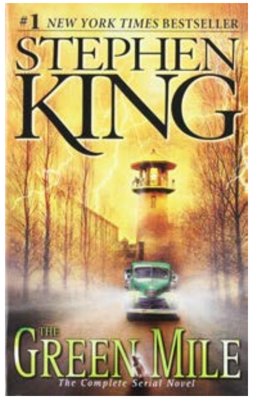
Single-Domain vs Cross-Domain

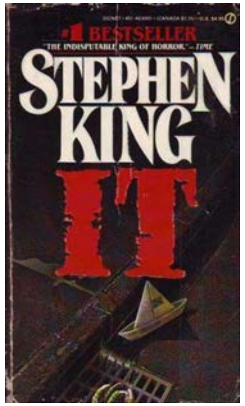
Single-Domain


COMMEMORATIVE EDITION
Pan
The Restaurant at the End of the Universe
DOUGLAS ADAMS
The sequel to
THE HITCH-HIKER'S GUIDE TO THE GALAXY


DON'T PANIC!
The Hitchhiker's Guide to the Galaxy
DOUGLAS ADAMS
AUTHOR OF THE RESTAURANT AT THE END OF THE UNIVERSE

Cross-Domain

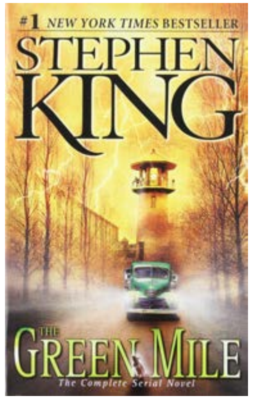


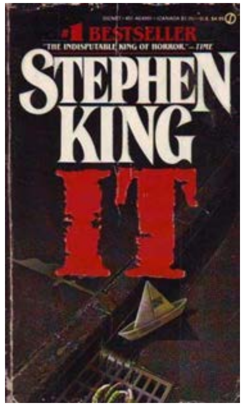


PAN 2019 Authorship Attribution Task

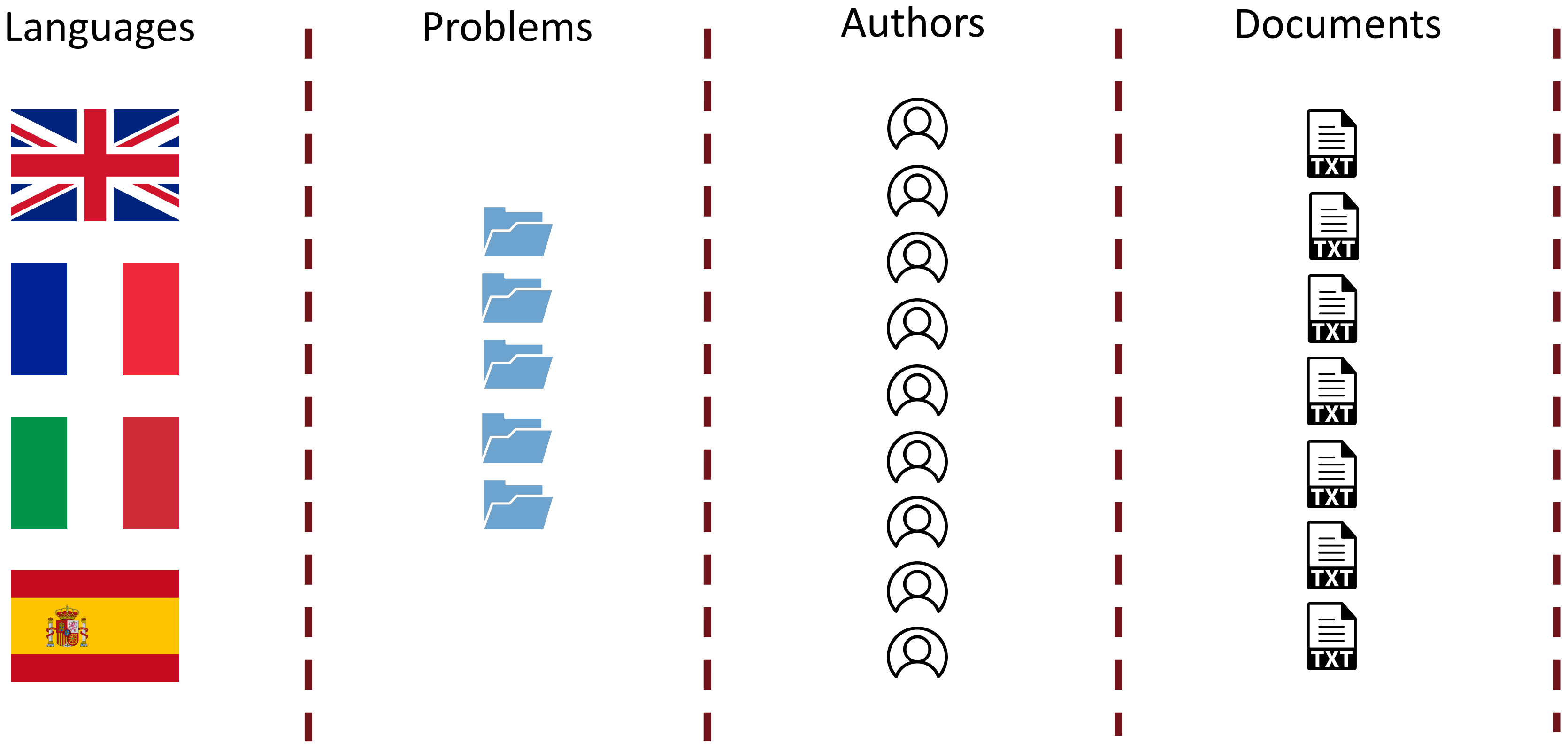
Open-set

Cross-Domain





PAN Dataset



Languages

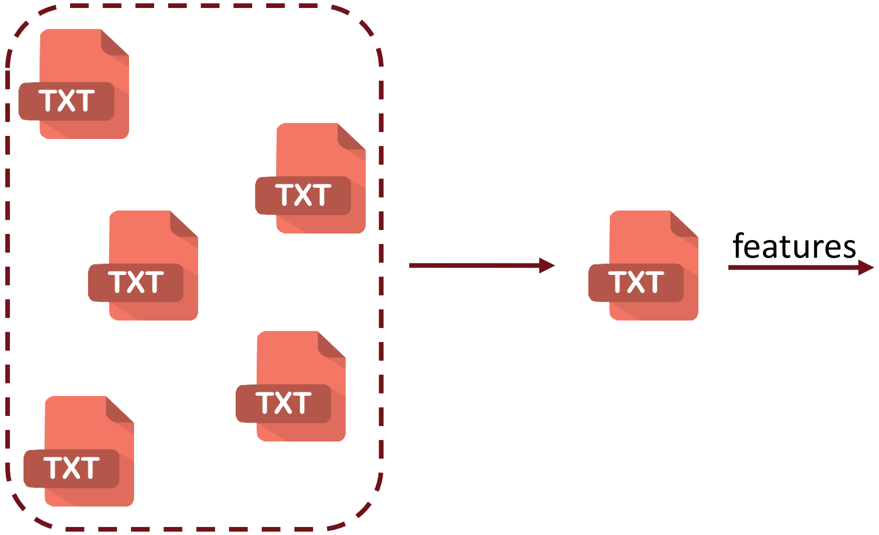
Problems

Authors

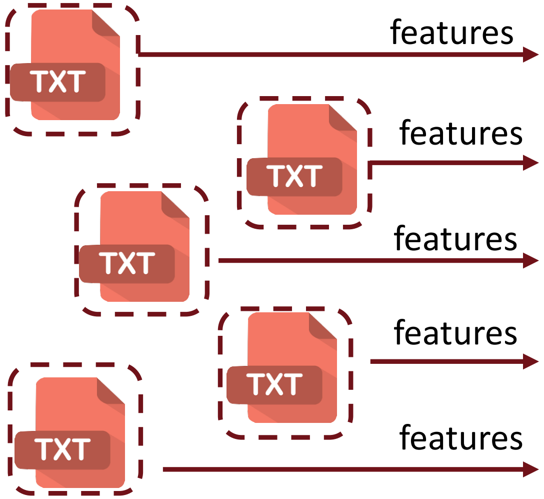
Documents

Main approaches to AA problems

Profile-Based Features

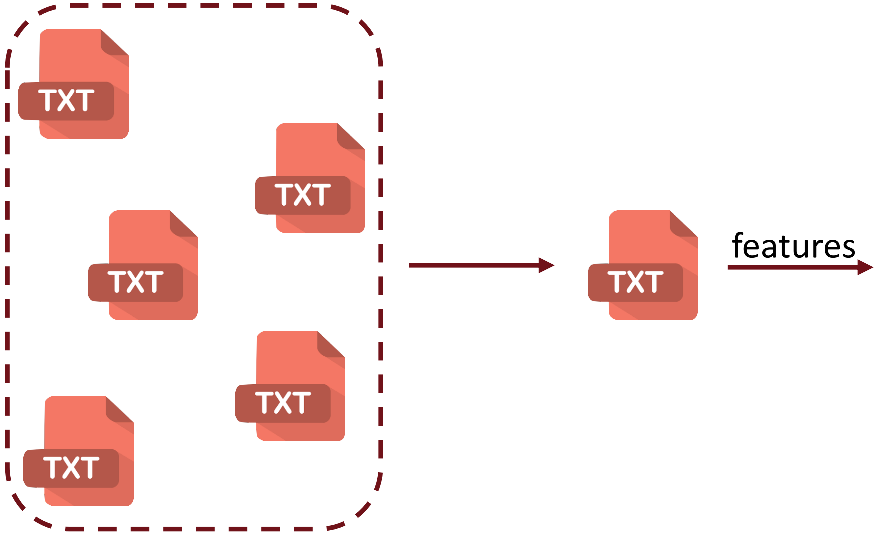


Instance-Based Features



Profile-Base features

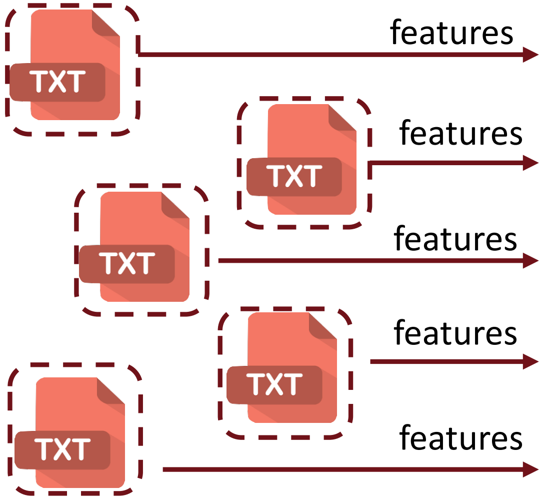
Profile-Based Features



- Concatenate together texts of the same author.
- Collecting as more information of the user as possible.
- Differences between the training texts by the same author are disregarded.
- Stylometric measures extracted from the concatenated file may be quite different in comparison to each of the original training texts.

Instance-Based Features

Instance-Based Features



- Analyze the texts associated with an author separately.
- Classification algorithms require multiple training instances per class for extracting a reliable model.
- The text samples should be long enough so that the text representation features can represent adequately their style.

Text Pre-Processing

- Pre-processing is a crucial step to prepare the data in almost every NLP problems.

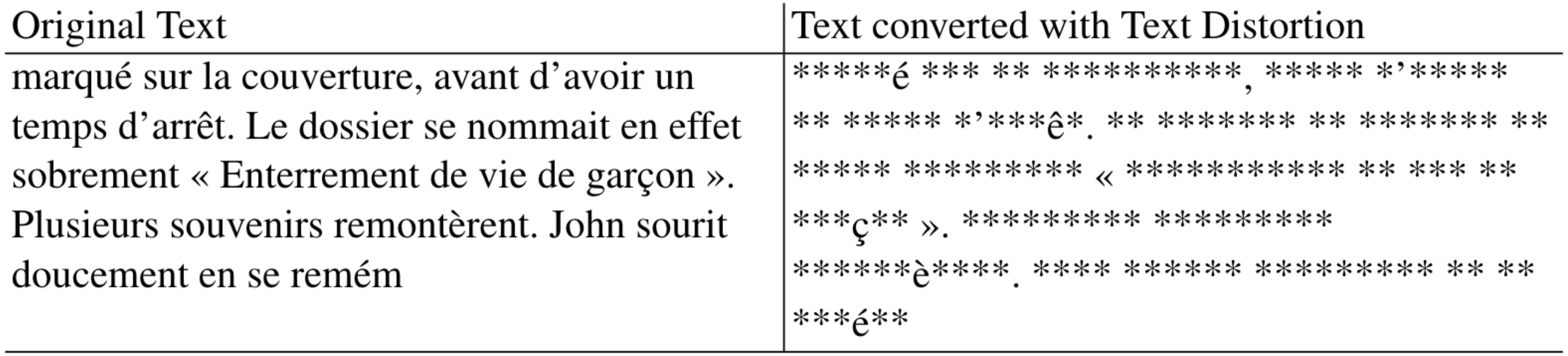
- Text pre-processing usually consists in normalize, sanitize or alter the text to remove noise, error, or completely change the data format.

- We used:

WordPunctTokenizer SnowballStemmer spaCy POS Tagger

Text Distortion

Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the 15° Conference of the European Chapter of the Association for Computational Linguistics:Volume 1, Long Papers. pp. 1138–1149 (2017)



Original Text	Text converted with Text Distortion
marqué sur la couverture, avant d’avoir un temps d’arrêt. Le dossier se nommait en effet sobrement « Enterrement de vie de garçon ». Plusieurs souvenirs remontèrent. John sourit doucement en se remém	*****é *** ** **********, ***** *’***** ** ***** *’***ê*. ** ******* ** ******* ** ***** ********* « *********** ** *** ** ***ç** ». ********* ********* ******è****. **** ****** ********* ** ** ***é**

SAPIENZA
UNIVERSITÀ DI ROMA

Features

Profile

Char3-5

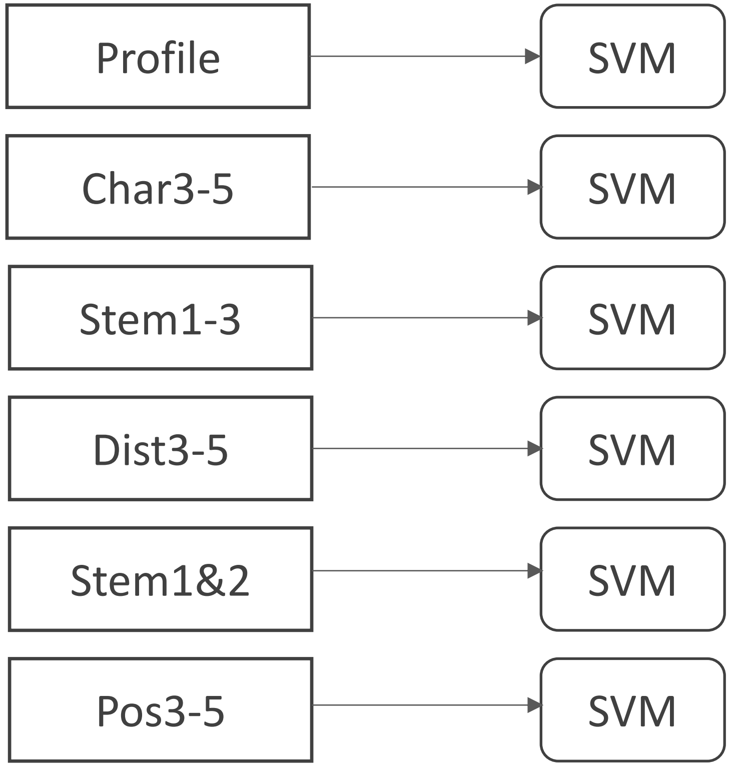
Stem1-3

Dist3-5

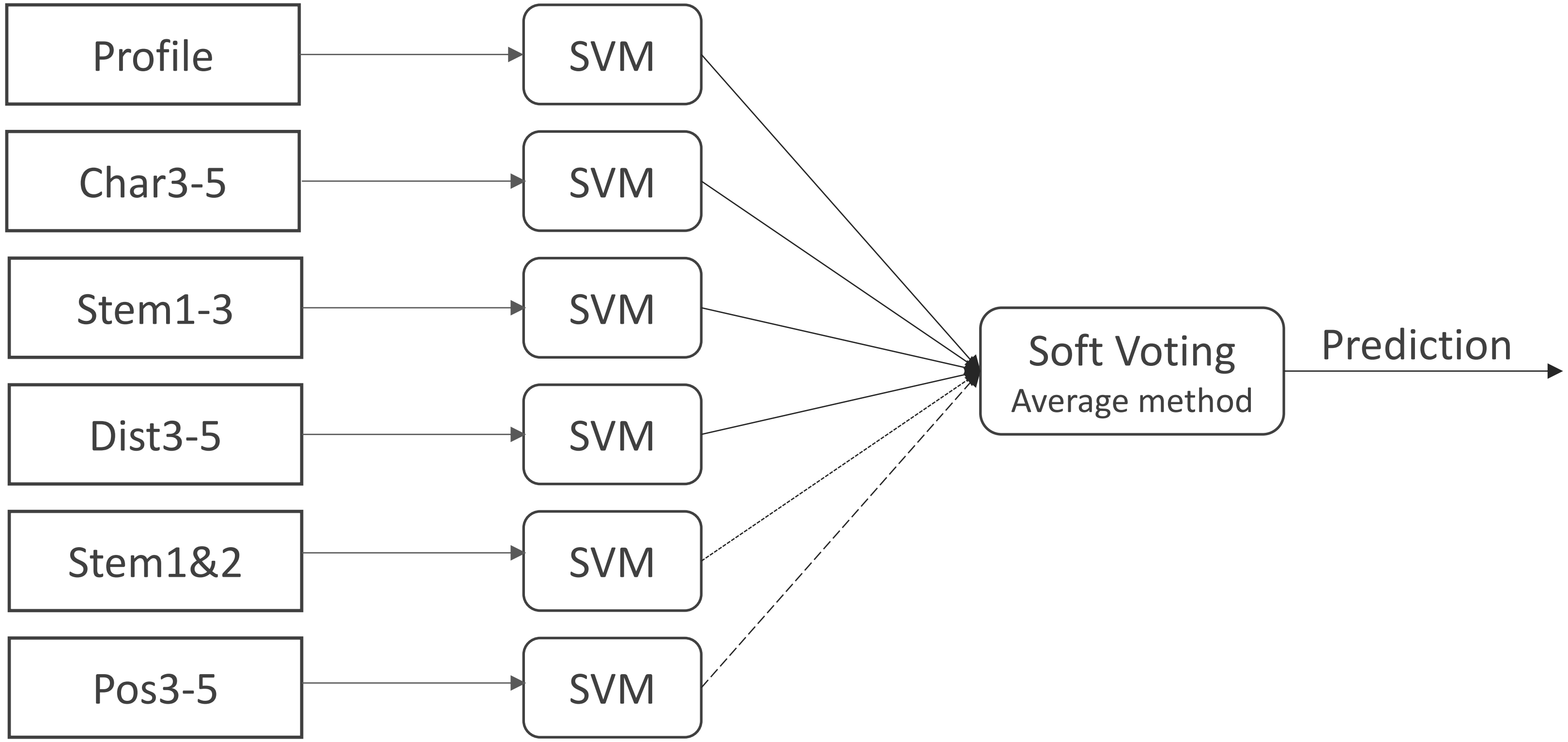
Stem1&2

Pos3-5

Model


Profile
SVM
Char3-5
SVM
Stem1-3
SVM
Dist3-5
SVM
Stem1&2
SVM
Pos3-5
SVM

Model

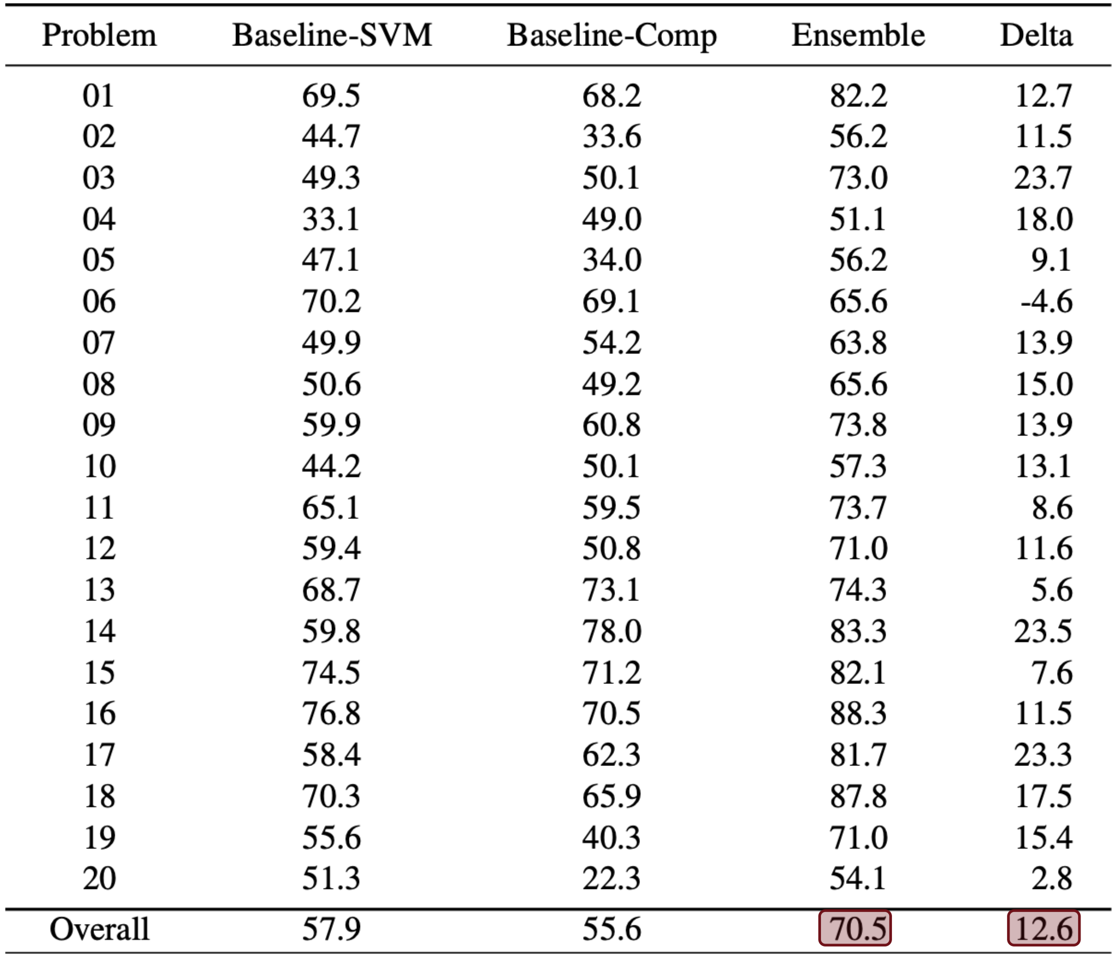

Profile
Char3-5
Stem1-3
Dist3-5
Stem1&2
Pos3-5
SVM
SVM
SVM
SVM
SVM
SVM
Soft Voting
Average method
Prediction

Unknown Prediction

P_i i-th most probable author for a given text

$$Unknown = \begin{cases} True, P_1 - P_2 < 0.1 \ \wedge mean\ (P_1 - \ P_2, P_1 - P_3) < 0.7 \\ False, otherwise \end{cases}$$

Result on DEV

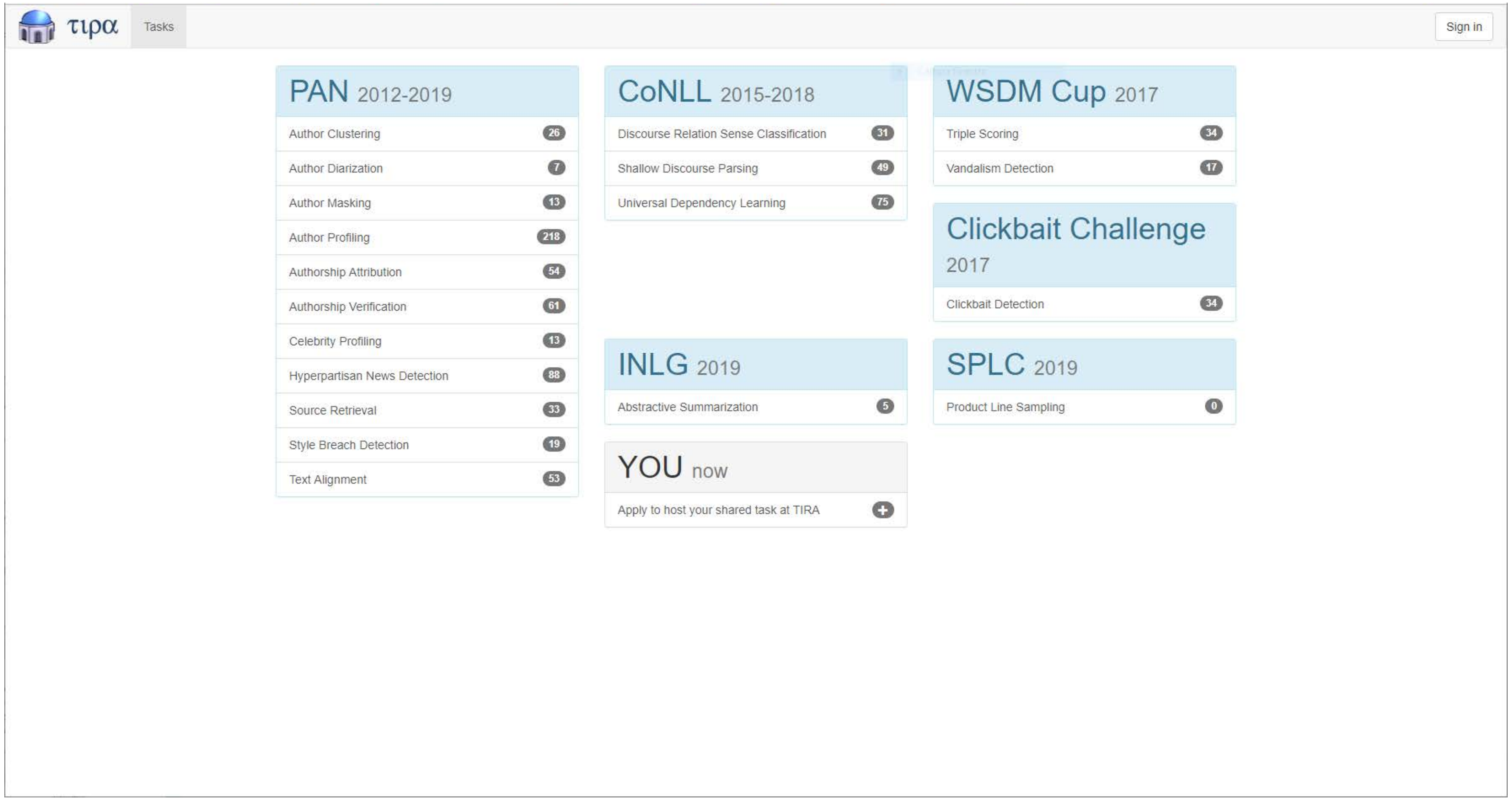


Problem	Baseline-SVM	Baseline-Comp	Ensemble	Delta
01	69.5	68.2	82.2	12.7
02	44.7	33.6	56.2	11.5
03	49.3	50.1	73.0	23.7
04	33.1	49.0	51.1	18.0
05	47.1	34.0	56.2	9.1
06	70.2	69.1	65.6	-4.6
07	49.9	54.2	63.8	13.9
08	50.6	49.2	65.6	15.0
09	59.9	60.8	73.8	13.9
10	44.2	50.1	57.3	13.1
11	65.1	59.5	73.7	8.6
12	59.4	50.8	71.0	11.6
13	68.7	73.1	74.3	5.6
14	59.8	78.0	83.3	23.5
15	74.5	71.2	82.1	7.6
16	76.8	70.5	88.3	11.5
17	58.4	62.3	81.7	23.3
18	70.3	65.9	87.8	17.5
19	55.6	40.3	71.0	15.4
20	51.3	22.3	54.1	2.8
Overall	57.9	55.6	70.5	12.6

Further analysis

- Closed-Set scenario accuracy of 87% on a total of 2,646 documents.
- Closed-Set scenario with Unknowns detector achieve as overall result an accuracy of 78.7%
- Difference in results of 8.7%

TIRA



PAN 2012-2019

Author Clustering	26
Author Diarization	7
Author Masking	13
Author Profiling	218
Authorship Attribution	54
Authorship Verification	61
Celebrity Profiling	13
Hyperpartisan News Detection	88
Source Retrieval	33
Style Breach Detection	19
Text Alignment	53

CoNLL 2015-2018

Discourse Relation Sense Classification	31
Shallow Discourse Parsing	49
Universal Dependency Learning	75

INLG 2019

Abstractive Summarization	5

YOU now

Apply to host your shared task at TIRA

WSDM Cup 2017

Triple Scoring	34
Vandalism Detection	17

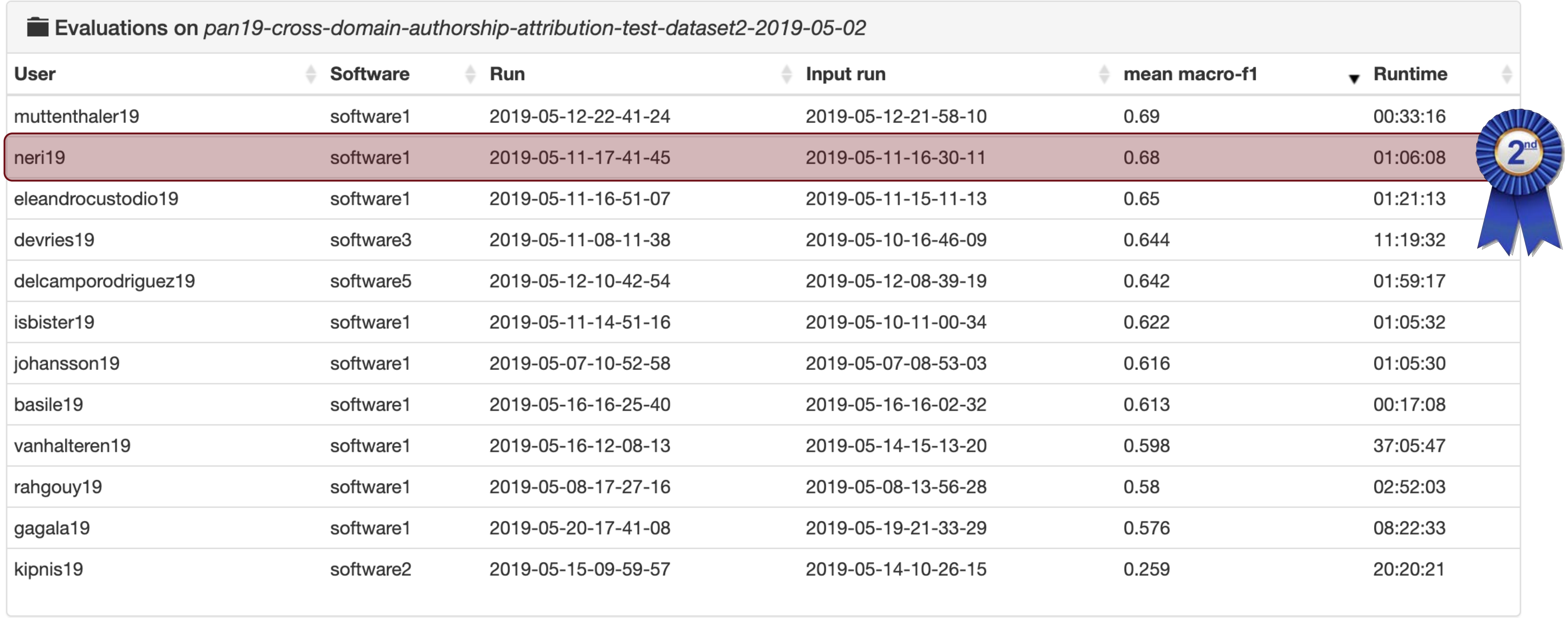
Clickbait Challenge 2017

Clickbait Detection	34

SPLC 2019

Product Line Sampling	0

Result



Evaluations on *pan19-cross-domain-authorship-attribution-test-dataset2-2019-05-02*

User	Software	Run	Input run	mean macro-f1	Runtime
muttenthaler19	software1	2019-05-12-22-41-24	2019-05-12-21-58-10	0.69	00:33:16
neri19	software1	2019-05-11-17-41-45	2019-05-11-16-30-11	0.68	01:06:08
eleandrocustodio19	software1	2019-05-11-16-51-07	2019-05-11-15-11-13	0.65	01:21:13
devries19	software3	2019-05-11-08-11-38	2019-05-10-16-46-09	0.644	11:19:32
delcamporodriguez19	software5	2019-05-12-10-42-54	2019-05-12-08-39-19	0.642	01:59:17
isbister19	software1	2019-05-11-14-51-16	2019-05-10-11-00-34	0.622	01:05:32
johansson19	software1	2019-05-07-10-52-58	2019-05-07-08-53-03	0.616	01:05:30
basile19	software1	2019-05-16-16-25-40	2019-05-16-16-02-32	0.613	00:17:08
vanhalteren19	software1	2019-05-16-12-08-13	2019-05-14-15-13-20	0.598	37:05:47
rahgouy19	software1	2019-05-08-17-27-16	2019-05-08-13-56-28	0.58	02:52:03
gagala19	software1	2019-05-20-17-41-08	2019-05-19-21-33-29	0.576	08:22:33
kipnis19	software2	2019-05-15-09-59-57	2019-05-14-10-26-15	0.259	20:20:21

Conclusion

- Ensemble model with a classifier for each feature.
- We combine Profile-Based and Instance-Based features together.
- We introduced a method that takes into account the three most similar author for the disputed text, instead of only the first two.
- We outperform the baseline in almost every problems.

Future Work

- Although our methodology to detect the unknown authors performs slightly better than the baseline, further improvements are needed.
- In one problem we reach a score lower than the baseline. It could be useful to understand the reason of it.
- Neural Networks approach could be tested.

THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
GRAZIE
MEHRBANI
GOZAIMASHITA
EFCHARISTO
KOMAPSUMNIDA
JUSPAXAR
PALDIES
MAAKE
BIYAN
SHUKRIA
TINGKI
BOLZÏN
MERCI

Question?