Special Topics in Text Mining

Manuel Montes y Gómez

http://ccc.inaoep.mx/~mmontesg/

mmontesg@inaoep.mx

Instituto Nacional de Astrofísica, Óptica y Electrónica

Style-based text classification

Agenda

- Introduction to style-based text classification
  - Tasks and applications
- Authorship attribution
  - Features and classification approaches
  - The Local Bag of words representation
- Author profiling
  - Features and classification approaches
  - A concise semantic representation for AP

Text classification

- It is the assignment of free-text documents to one or more predefined categories based on their content
- Important to remember:
  - Assigns documents to known categories
    - It does not aim to discover topics or classes
  - It is a supervised task: training data is required

But, can we only classified documents by their topic?

Text classification criteria

- Topic
  - Filtering of newswire stories
  - Indexing of scientific articles
  - Spam filtering
- Opinion
  - Sentiment analysis
- Style
  - Authorship analysis
  - Genre classification

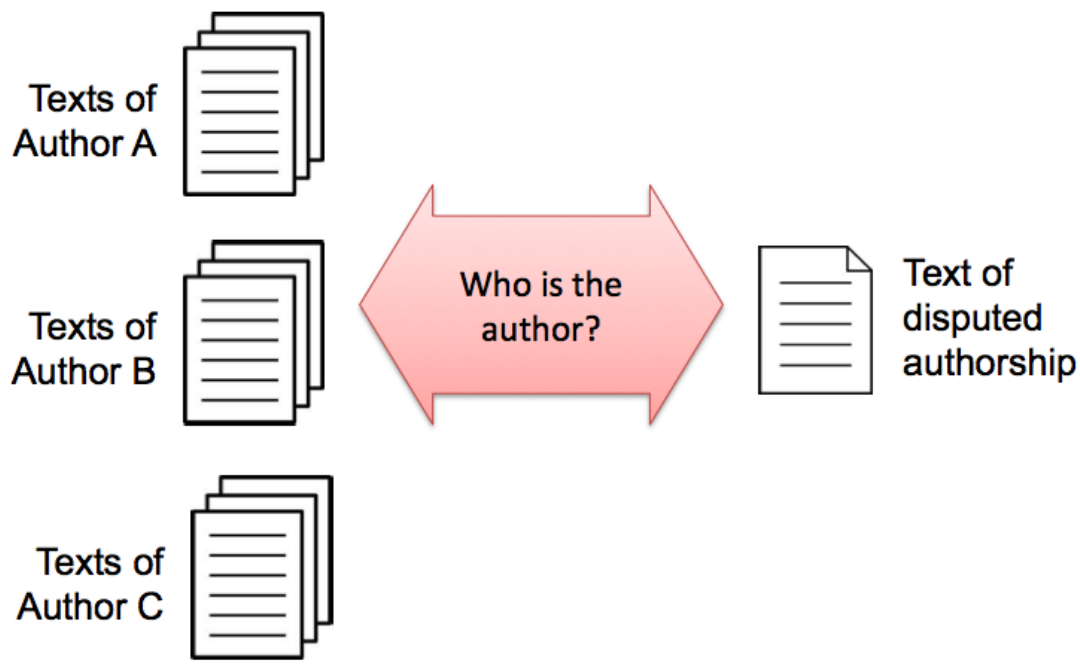
What kind of tasks?

Stamatatos, E. 2009. A Survey of Modern Authorship Attribution Methods, Journal of the American Society for information Science and Technology, 60(3): 538-556.
Stamatatos, E. 2015. An Introduction to Authoship Analysis. Material from tutorial at the 1st Mexican Autumn School on Language Technologies. Puebla, Mexico, Oct 2015.

Authorship analysis tasks

Authorship attribution

- Given a set of candidate authors and some texts by them, to attribute an unseen text to one of them.



Authorship analysis tasks

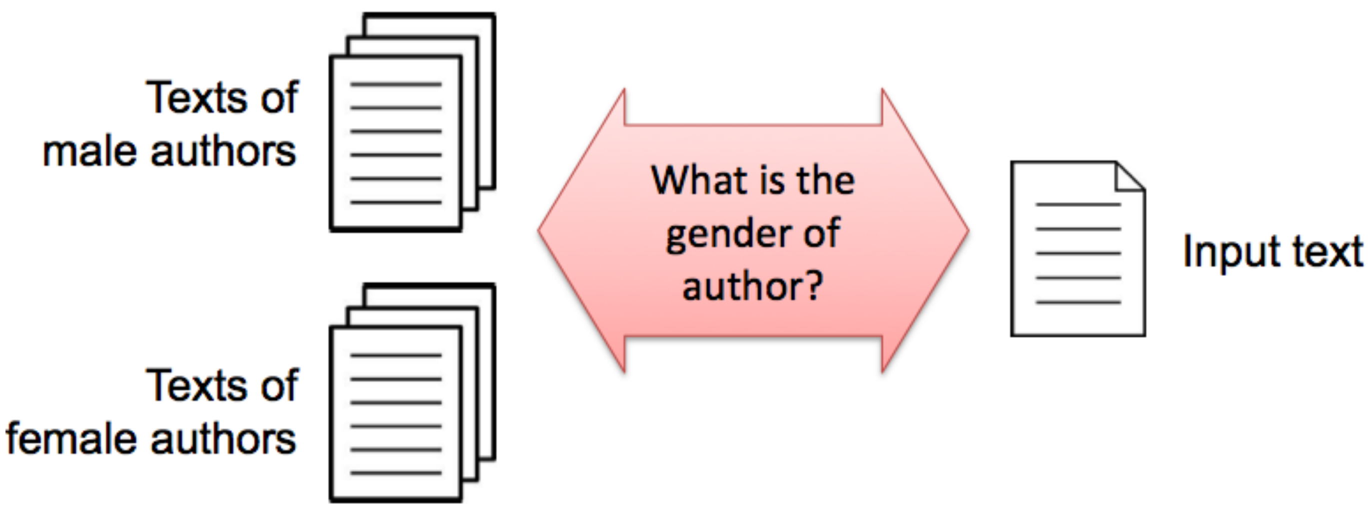
Author verification

- Given texts of a certain author, to decide whether an unseen text was written by that author or not.

Authorship analysis tasks

Author profiling

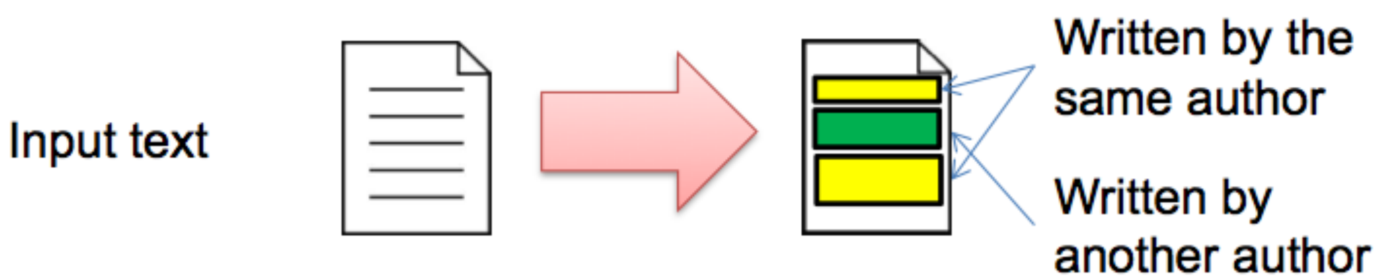
- Extraction of information about the age, gender, educational level, dialect, personality, etc. of the author.



Authorship analysis tasks

Author diarization

- Decompose a multi-author document into authorial components



Applications of these authorship analysis tasks?

Applications areas of AA

- Intelligence
  - Attribution of messages or proclamations to known terrorists
  - Linking different messages by authorship
- Criminal law
  - Identifying writers of harassing messages
  - Verifying the authenticity of suicide notes
- Civil law
  - Copyright disputes
- Computer forensics
  - Identifying the authors of source code of malicious software

Applications areas of AA (2)

- Literary research
  - Attributing anonymous or disputed literary works to known authors
  - Studying the differences among literary periods, schools, writers
- Historical research
  - Studying the writing style of an author (politician) in time
- Decision making
  - Personalized product advertisement

Authorship attribution

Stamatatos, E. 2009. A Survey of Modern Authorship Attribution Methods, Journal of the American Society for information Science and Technology, 60(3): 538-556.

AA as a classification problem

- In the typical authorship attribution problem, a text of unknown authorship is assigned to one candidate author, given a set of candidate authors for whom text samples of undisputed authorship are available.
- From a machine learning point-of-view, this can be viewed as a multi-class single-label text categorization task.

Is the BoW representation adequate for this task?

Features and methods

- The main idea behind AA is that by measuring some *textual features* we can distinguish between texts written by different authors.
- Important to have features that quantify the *writing style* of authors, and apply methods able to learn from that kind of features.

How to address the AA problem?

What features could be used?

Lexical features (1)

- Several different lexical features have been used in the task of AA:
  - Simple measures such as sentence length counts and word length counts
    - Can be applied to any language and any corpus
    - For certain languages is not trivial to do word segmentation → Chinese, German, etc.
  - Vocabulary richness and the number of hapax legomena (i.e., words occurring once).
    - Vocabulary size heavily depends on text-length

Lexical features (2)

- Traditional bag-of-words text representation
  - Good for topic classification, but not necessarily capture the writing style of authors.
- Function words
  - Are used in a largely unconscious manner by the authors and they are topic-independent
- Subset of more frequent words
  - Similar problems than bag-of-words
- Word n-grams
  - Not always better than individual word features
  - Dimensionality increases considerably

Character features

- According to this family of measures, a text is viewed as a mere sequence of characters.
- Various character-level measures:
  - alphabetic characters count, digit characters count, uppercase and lowercase characters count, letter frequencies, punctuation marks count, etc.
- Frequencies of character n-grams
  - Lexical information (e.g., |_in_|, |text|)
  - Contextual information (e.g., |in_t|)
  - Use of punctuation and capitalization
  - Common used suffix (e.g., |ful_|, |ing_|)

Syntactic features (1)

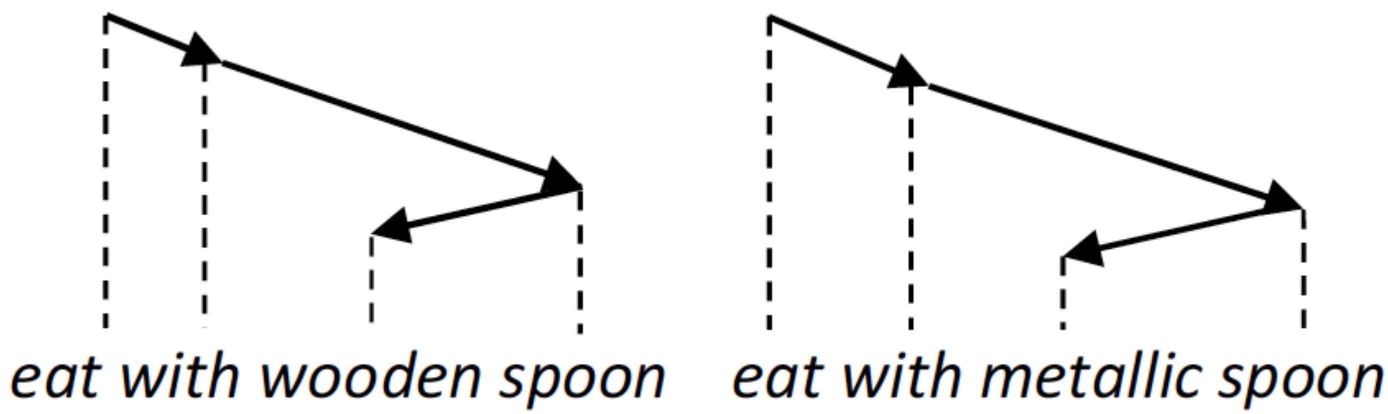
- The idea is that authors tend to use similar syntactic patterns unconsciously.
- Syntactic information is considered more reliable authorial *fingerprint* in comparison to lexical Information
- Disadvantages:
  - Robust and accurate NLP tools are require to perform syntactic analysis of texts
  - Language-dependent procedure

Syntactic features (2)

- POS tag frequencies or POS tag n-gram frequencies
  - A_DD few_JJ examples_NNS of_PREP heterologous_JJ expression_NN
- Noun phrase counts, verb phrase counts, length of noun phrases, length of verb phrases, etc.
  - NP[Another attempt] VP[to exploit] NP[syntactic information] VP[was proposed] PP[by Stamatatos, et al. (2000)].

Syntactic features (3): recent approaches

- Using syntactic-based n-grams as features
  - Sn-grams are obtained based on the order in which the elements are presented in syntactic trees.

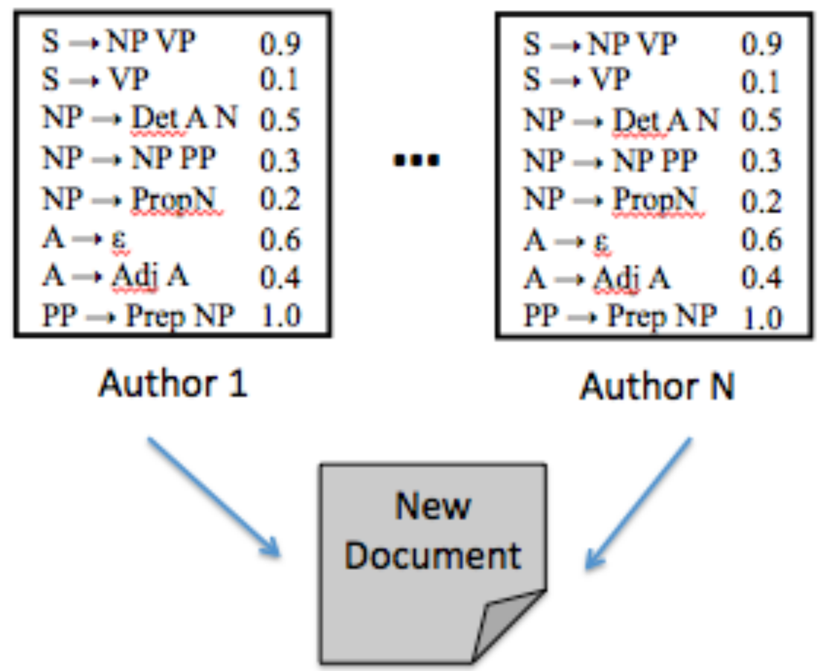


eat with, **with spoon, eat with spoon**

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, Liliana Chanona-Hernández. *Syntactic Dependency-Based N-grams as Classification Features*. Lecture Notes in Computer Science Volume 7630, 2013.

Syntactic features (4): recent approaches

- Using probabilistic context free grammars as language models for classification
  - Generate a parse tree for each training document
  - Estimate a grammar and its parameters from the assembled “tree-bank”.
  - Compute probabilities for each document, for each grammar
  - Select the author (grammar) with the highest probability



$$P(Doc \mid \text{Author 1}) > P(Doc \mid \text{Author N})\ ?$$

Semantic features

- The more detailed the text analysis required for extracting features, the less accurate the produced measures.
  - Few attempts to exploit high-level features
- Examples of the usage of semantic information:
  - Use semantic relations (from dependency trees)
  - Use synonyms and hypernyms of words (Wordnet)
  - Detect semantic similarity between words by means of LSI

Domain-specific features

- In some applications it is possible to use some structural measures to quantify the authorial style.
- Some examples are:
  - Use of greetings and farewells in the messages
  - Types of signatures
  - Use of indentation
  - Paragraph length
  - Font color counts and font size counts

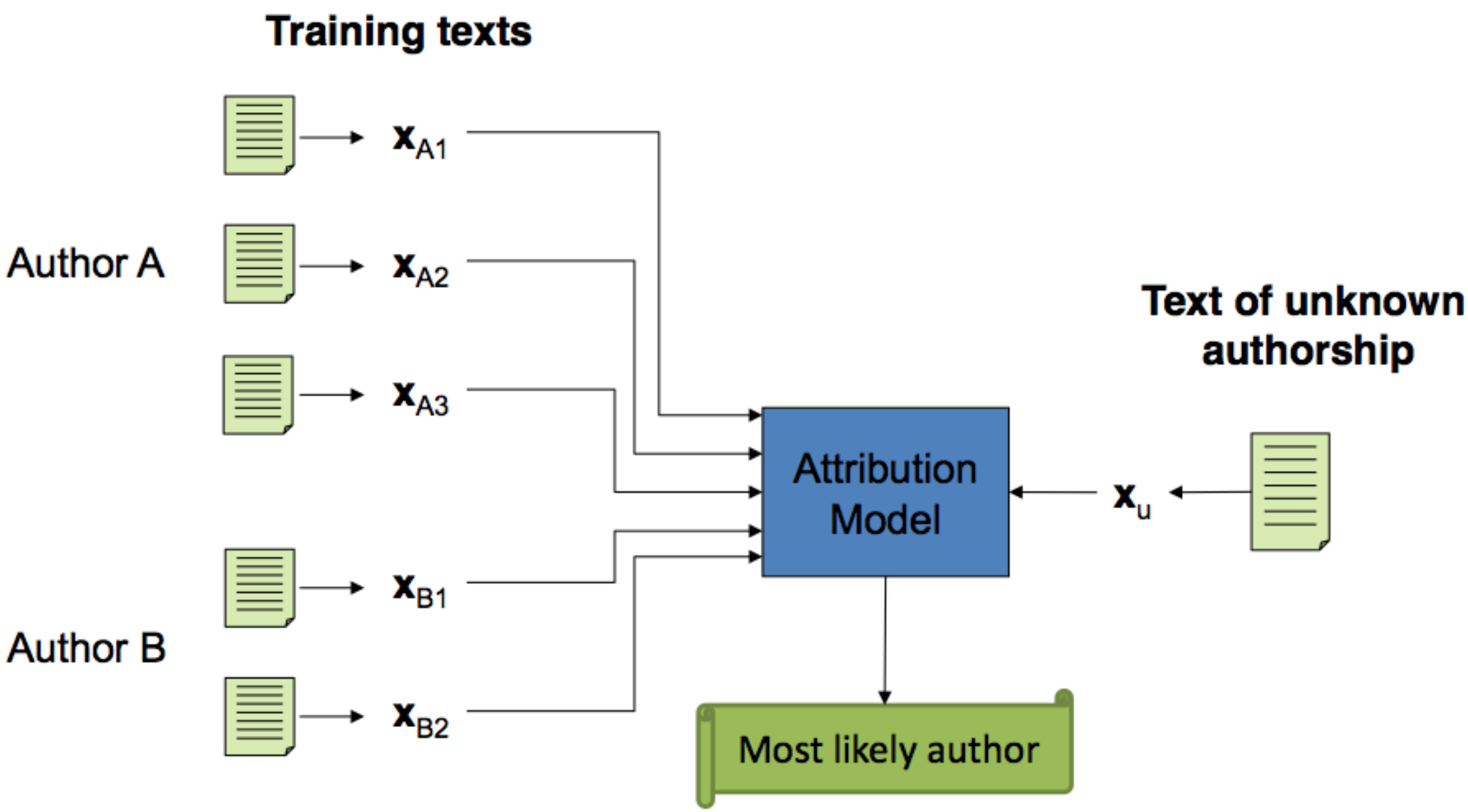
Authorship attribution methods

- Instance-based approaches
  - Each training text is individually represented as a separate instance of authorial style.
  - Uses vector space representations and apply supervised learning algorithms such as traditional text classification.
- Profile-based approaches
  - Concatenate all the available training texts per author in one big file and extract a cumulative representation of that author's style (profile) from this concatenated text.

What is better?

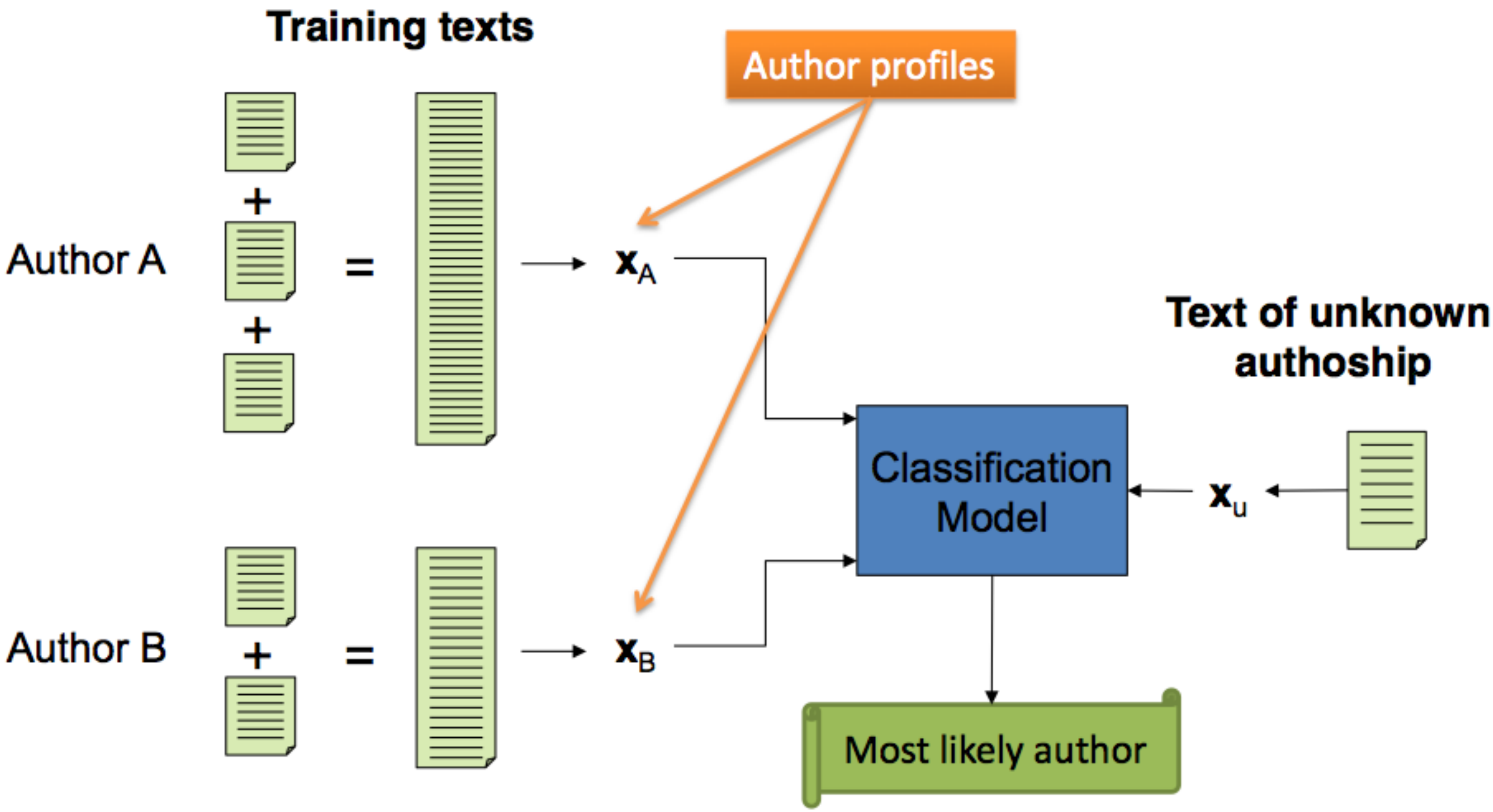
Advantages and disadvantages?

Instance based approach



Stamatatos, E. 2015. An Introduction to Authoship Analysis. Material from tutorial at the 1st Mexican Autumn School on Language Technologies. Puebla, Mexico, Oct 2015.

Profile based approach



Stamatatos, E. 2015. An Introduction to Authoship Analysis. Material from tutorial at the 1st Mexican Autumn School on Language Technologies. Puebla, Mexico, Oct 2015.

Profile-based approaches (1)

- Training just comprises the extraction of profiles for the candidate authors.
- Attribution is based on the distance of the profile of an unseen text and the profile of each author.

$$author(x) = \arg\min_{a \in \mathbf{A}} d(PR(x), PR(x_a))$$

- It can be realized by using probabilistic and compression models

Profile-based approaches (2)

- Probabilistic models: attempt to maximize the probability $P(x|a)$ *for a text x to* belong to an author *a.*
  - Can be applied to both character and word sequences

$$author(x) = \arg\max_{a \in \mathbf{A}} \log_2 \frac{P(x \mid a)}{P(x \mid \bar{a})}$$

- Compression models: the difference in bit-wise size of the compressed files $d(x, x_a)=C(x_a+x)-C(x_a)$ indicates the similarity of text *x* with author *a*.
  - Several compression algorithms have been tested including RAR, LZW, GZIP, BZIP2, 7ZIP.

Some comments on AA

- The number of candidate authors
  - Increasing the number of authors leads to a significant decrease in performance
  - Character n-grams outperform other feature types
- The size of the training set
  - AA can lead to reasonable results even when only limited data is available
  - Character n-grams show more robustness to the effect of data size than syntactic or word-based features
- The instance-based approach usually reports better results than the profile-based approach

Our proposal: using the LOWBOW representation

- BOW shows acceptable performance, particularly using word and character n-grams features.
  - It only takes into account the occurrence of n-grams
- BOW ignores any sequential information in documents
- Our proposal is to use richer document representations for AA that incorporate sequential information.
  - The distribution of terms at different locations can reveal useful (stylistic) information about authors

H. J. Escalante, T. Solorio, M. Montes. *Local Histograms of Character Ngrams for Authorship Attribution*. ACL Conference. Portland, Oregon, June 20, 2011

The traditional BOW approach

- Indicates the (weighted) occurrence of terms in a document

From Nov 10, 1999
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

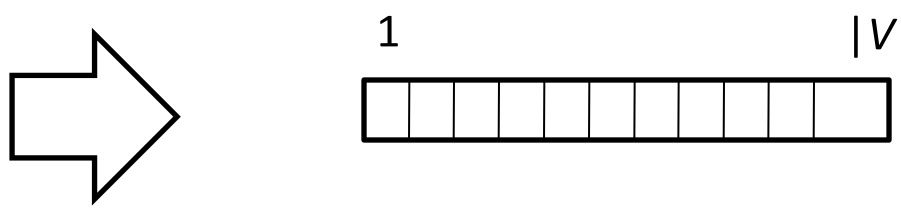
We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim



$$\mathbf{d}_i = \left[x_{i,1}, \ldots, x_{i,|V|} \right]$$

The LOWBOW framework

- Consider a set of histograms, each weighted according to selected positions in the document

From Nov 10, 1999
Xm Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

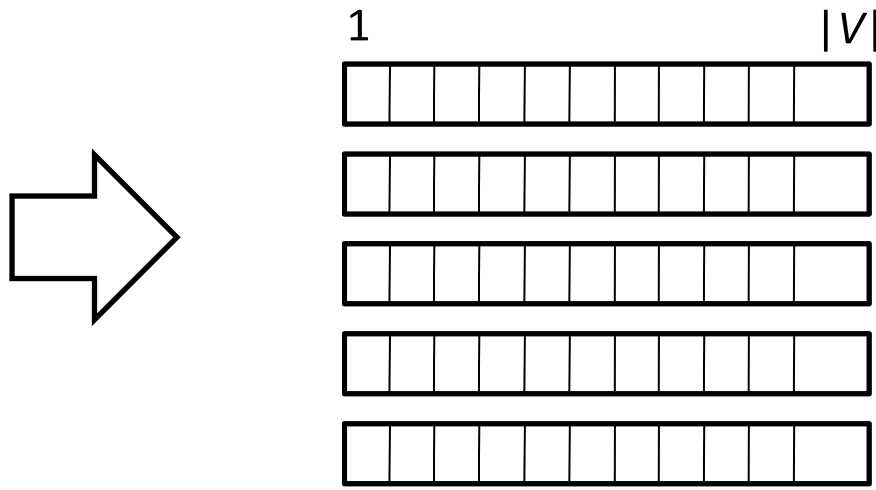
We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Xm



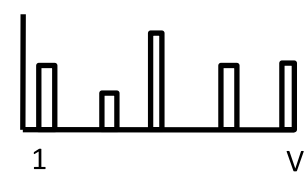
$$\mathbf{d}_i = \{\mathbf{dl}_i^1, \ldots, \mathbf{dl}_i^k\} \qquad \mathbf{dl}_i^j = \mathbf{d}_i^G \times K_{\mu_j,\sigma}^s \qquad \mathbf{d}_i^G = \left[x_{i,1}, \ldots, x_{i,|V|}\right]$$

BOW computation

China sent a senior official to attend a reception at the Ukraine embassy on Friday despite a diplomatic rift over a visit to Kiev by Taiwan's vice president Lien Chan. But an apparent guest list mix-up left both sides unsure over who would represent Beijing at the reception, held to mark Ukraine's independence day…

Benjamin Kang Lim

BOW representation



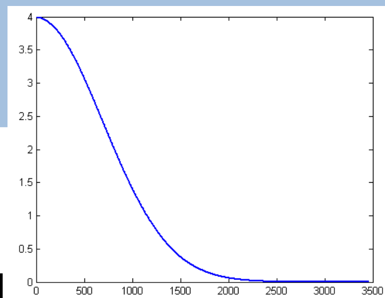
LOWBOW computation

China sent a senior official to attend a reception at the Ukraine embassy on Friday despite a diplomatic rift over a visit to Kiev by Taiwan's vice president Lien Chan. But an apparent guest list mix-up left both sides unsure over who would represent Beijing at the reception, held to mark Ukraine's independence day…

Benjamin Kang Lim

Identify locations in documents

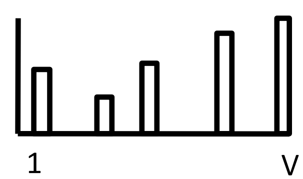
BOW computation



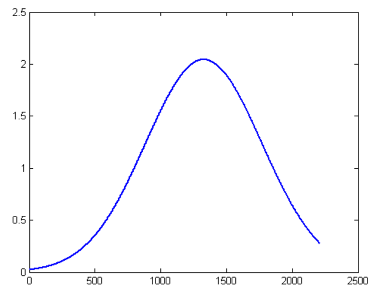
China sent a senior official to attend a reception at the Ukraine embassy on Friday despite a diplomatic rift over a visit to Kiev by Taiwan's vice president Lien Chan. But an apparent guest list mix-up left both sides unsure over who would represent Beijing at the reception, held to mark Ukraine's independence day…

Benjamin Kang Lim

Weight the contribution of terms according to Gaussians at the different locations



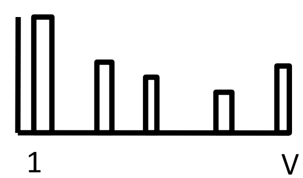
LOWBOW computation



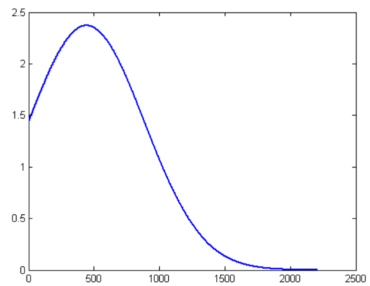
Chi or official to attend a reception at the Ukr on Friday despite a diplomatic rift over a vi Taiwan's vice president Lien Chan. But an apparent guest list mix-up left both sides unsure over who would represent Beijing at the reception, held to mark Ukraine's independence day…

Benjamin Kang Lim

Weight the contribution of terms according to Gaussians at the different locations



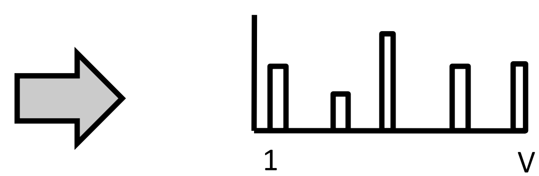
LOWBOW



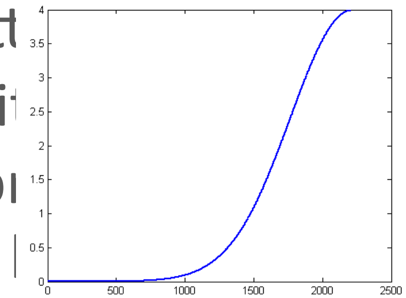
China sent ial to attend a reception at the Ukraine er ay despite a diplomatic rift over a visit to K 's vice president Lien Chan. But an apparent guest list mix-up left both sides unsure over who would represent Beijing at the reception, held to mark Ukraine's independence day…

Benjamin Kang Lim

Weight the contribution of terms according to Gaussians at the different locations

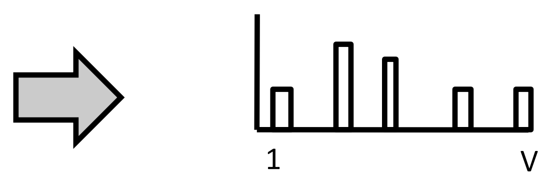


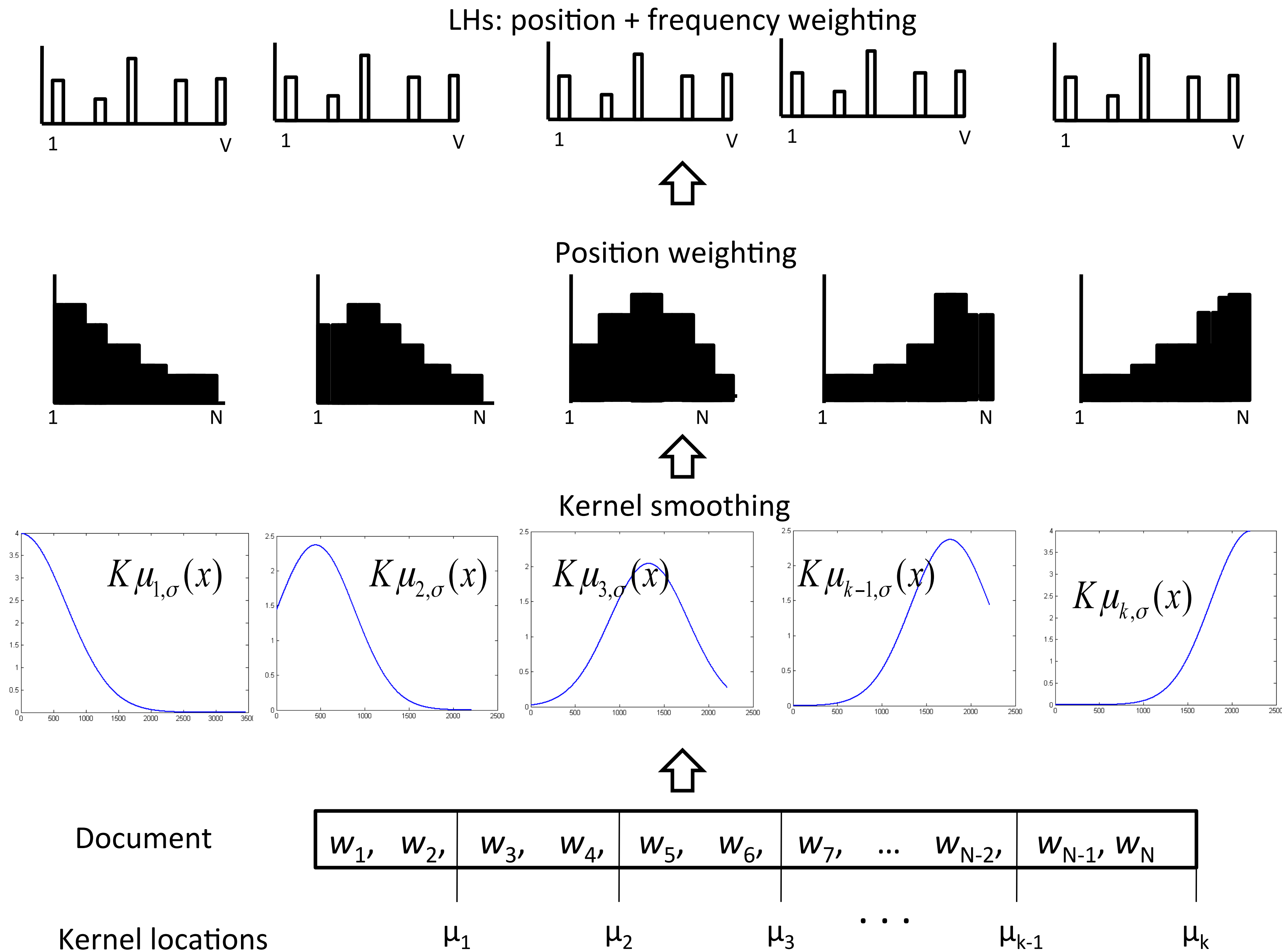
LOWBOW

China sent a senior official to att  n at the Ukraine embassy on Friday despi rift over a visit to Kiev by Taiwan's vice pr ıan. But an apparent guest list mix-up unsure over who would represent Beijing at the reception, held to mark Ukraine's independence day…

Benjamin Kang Lim

Weight the contribution of terms according to Gaussians at the different locations





LHs: position + frequency weighting
1
V
Position weighting
1
N
Kernel smoothing
$K\mu_{1,\sigma}(x)$
$K\mu_{2,\sigma}(x)$
$K\mu_{3,\sigma}(x)$
$K\mu_{k-1,\sigma}(x)$
$K\mu_{k,\sigma}(x)$
Document
w_1, w_2, w_3, w_4, w_5, w_6, w_7, … w_{N-2}, w_{N-1}, w_N
Kernel locations
μ_1
μ_2
μ_3
…
μ_{k-1}
μ_k

AA using LOWBOW

- As classifier we used SVM; we needed a new way to measure de distance between documents:
  - Combining vectors in one single vector.
    $$\mathbf{L}_i = \sum_{j=1}^{k} \mathbf{dl}_i^j$$
  - Representing documents by a set of vectors.
    $$\mathbf{L}_i = \{\mathbf{dl}_i^1, \ldots, \mathbf{dl}_i^k\}$$
- We used the Euclidean and Chi-squere distances.

$$D(P,Q) = \sum_{l=1}^{k} \sum_{i=1}^{|V|} \sqrt{(\mathbf{p}_l^i - \mathbf{p}_l^i)^2} \qquad D(P,Q) = \sum_{l=1}^{k} \sum_{i=1}^{|V|} \frac{(\mathbf{p}_l^i - \mathbf{q}_l^i)^2}{(\mathbf{p}_l^i + \mathbf{q}_l^i)}$$

Experimental settings

- We consider a subset of RCV-I, documents written by *10 authors* (about the same subject); 50 documents are available for training and 50 for testing for each author
- Experiments using *words* and *3-grams at the character* level were performed, different number of locations and scale parameters were evaluated, we report the settings that showed better performance
- The 2500 most frequent terms were used to obtain the representations

S. Plakias and E. Stamatatos. **Tensor space models for authorship attribution.** *LNCS 5138,* pp. 239–249, Springer, 2008.

Results using 50 training documents per author

Method	Parameters	Words	Char. N-grams
BOW	-	78.2%	75.0%
1-vector	k = 2; σ = 0.2	75.8%	72.0%
1-vector	k = 5; σ = 0.2	77.4%	75.2%
1-vector	k = 20; σ = 0.2	77.4%	75.0%

BOW is a strong baseline

k	Euc.	Chi^2
Words		
2	78.6%	75.4%
5	77.6%	77.2%
20	79.2%	79.0%
Character N-grams		
2	83.4%	83.8%
5	83.4%	84.6%
20	84.6%	85.2%

K-vectores

The best accuracy using character n-grams and the chi-squere distance

Results using smaller training sets

- Using words as features

Método \ conjunto	1-doc	3-docs	5-docs	10-docs	50-docs
BOW	36.8%	57.1%	62.4%	69.9%	78.2%
1-vector	37.9%	55.6%	60.5%	69.3%	77.4%
K-vectores	52.4%	63.3%	69.2%	72.8%	82.0%
Referencia	-	-	53.4%	67.8%	80.8%

- Using character n-grams as features

Método \ conjunto	1-doc	3-docs	5-docs	10-docs	50-docs
BOW	65.3%	71.9%	74.2%	76.2%	75.0%
1-vector	61.9%	71.6%	74.5%	73.8%	75.0%
K-vectores	70.7%	78.3%	80.6%	82.2%	86.4%
Reference	-	-	53.4%	67.8%	80.8%

Author profiling

The author profiling task

- It consists in knowing as much as possible about an unknown author, just by analyzing a given text.
  – Age, gender, social/economic status, level of studies, nationality, religion, etc.

- Some applications have to do with business intelligence, computer forensics and security.

Author profiling – main approach

- It is commonly approached as a single-label multiclass classification problem, where profiles represent the classes to discriminate.
- It involves three tasks:
  1. The extraction of features (words, style markers, etc.)
  2. The representation of documents
  3. The use of a machine learning method for inducing a classification model.

Which features?
The same than for authorship attribution?
For this task, what is more important, content or style?

Who wrote these reviews?

- Male or female?
  Mexican, Argentin or Spanish?

La ubicación no es en una linda zona de Lima, creo que no es una opción para hacer turismo en Lima. Es excelente para una estadía en ocasión de tránsito ya que está cerca del aeropuerto y el servicio de transfer es muy bueno.

La ubicación es excelente.

Entrar y un olor a cebolla que era insoportable. Había una delegación de malasia distribuyendo comida mucha gente casi a la entrada del hotel. No deberían permitir esto.ya que se impregno todo hasta los ascensores. Tampoco me gusto, que pusieran en mi tarjeta un monto superior al pactado con booking que presente el dia que llegue, pero a la salida en mi tarjeta había un monto superior. Puse mi queja y les mostre nuevamente el boucher y lo hicieron de nuevo. En fin hay que fijarse bien.

Nada. Una vergüenza de hotel.

Chalet en barrio perdido que han mal acondicionado como pretendido hotel. Muy ruidoso: se oye al resto de huéspedes, la calle, los aviones... En el chalet de al lado había una fiesta y la música sonaba atronadora, no pudimos descansar en toda la noche. Los jóvenes de la recepción son unos pusilánimes que no dan ninguna solución y a las 4 de la mañana, cuando nos fuimos al aeropuerto seguía la fiesta.

Who worte these reviews?

- Male or female?
  Mexican, Argentin or Spanish?

Edgardo
Grupo de amigos
Argentina

La ubicación no es en una linda zona de Lima, creo que no es una opción para hacer turismo en Lima. Es excelente para una estadía en ocasión de tránsito ya que está cerca del aeropuerto y el servicio de transfer es muy bueno.

Graciela
Persona que viaja sola
Argentina
21 de octubre de 2013

La ubicación es excelente.

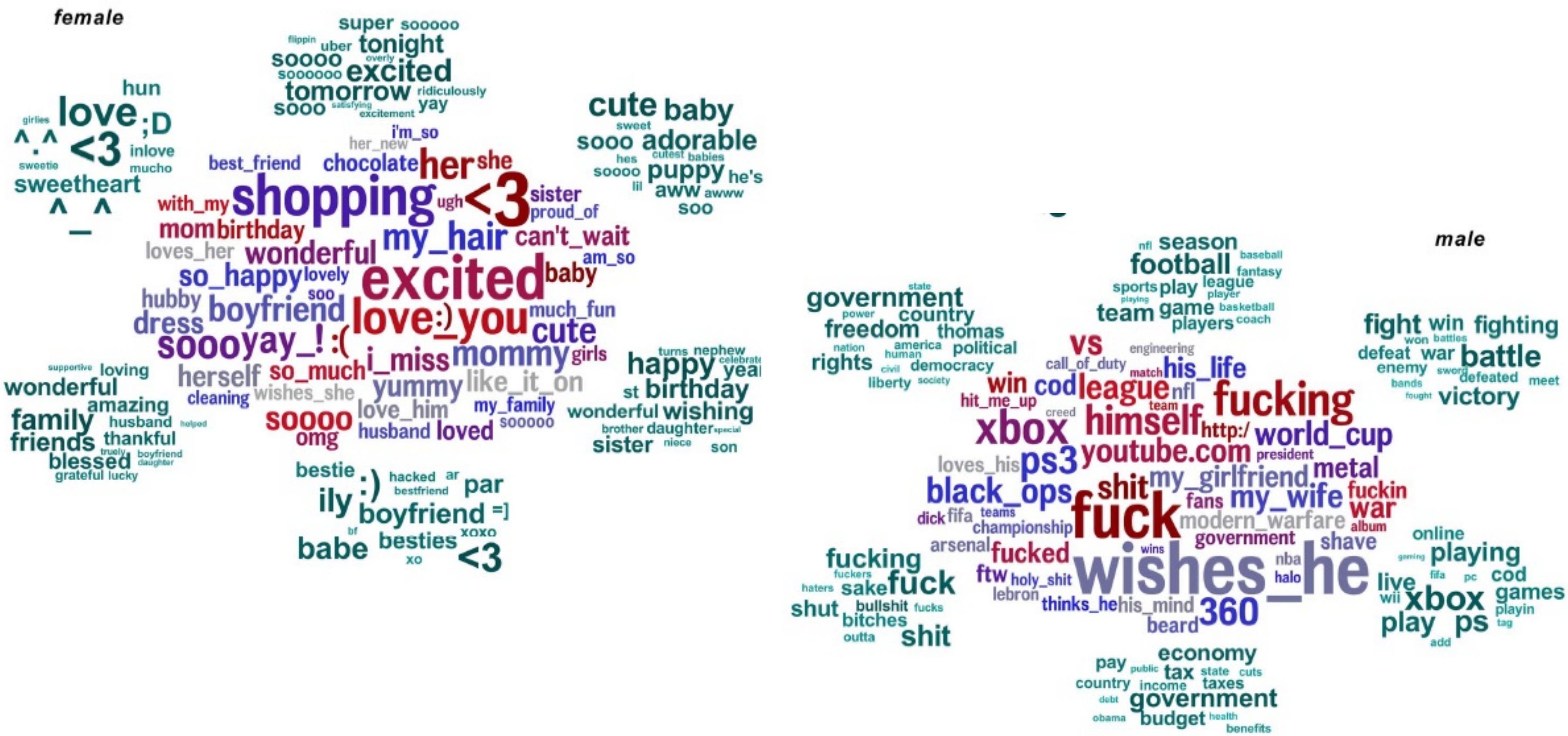
Entrar y un olor a cebolla que era insoportable. Había una delegación de malasia distribuyendo comida mucha gente casi a la entrada del hotel. No deberían permitir esto.ya que se impregno todo hasta los ascensores. Tampoco me gusto, que pusieran en mi tarjeta un monto superior al pactado con booking que presente el dia que llegue, pero a la salida en mi tarjeta había un monto superior. Puse mi queja y les mostre nuevamente el boucher y lo hicieron de nuevo. En fin hay que fijarse bien.

Laura
Pareja
España
15 de enero de 2014

Nada. Una vergüenza de hotel.

Chalet en barrio perdido que han mal acondicionado como pretendido hotel. Muy ruidoso: se oye al resto de huéspedes, la calle, los aviones... En el chalet de al lado había una fiesta y la música sonaba atronadora, no pudimos descansar en toda la noche. Los jóvenes de la recepción son unos pusilánimes que no dan ninguna solución y a las 4 de la mañana, cuando nos fuimos al aeropuerto seguía la fiesta.

Frequent words by women and men



• Schwartz et al. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791.

Author profiling – representation

- The most common approach is the Bag-of-Features (words, POS tags, emoticons, etc.)
- Some shortcomings of this representation are:
  - It produce high dimensionality and dispersion of information.
  - It does not preserve any kind of relationship of terms.

The English corpus used at PAN 2013:
• 236,000 instances, each instance is a text le with multiple blogs/posts by the same author.
• A total of 413,564 blogs/posts and 180,809,187 words (*more than 5 million different "words"*)

Our idea in a few words

- Use very simple but highly effective meta-attributes for representing the documents
- Our intention is to reduce the dimensionality problem, on one hand, and to capture the relation between words and between words and profiles
- These attributes were inspired in some ideas from distributional representations and concise semantic analysis.

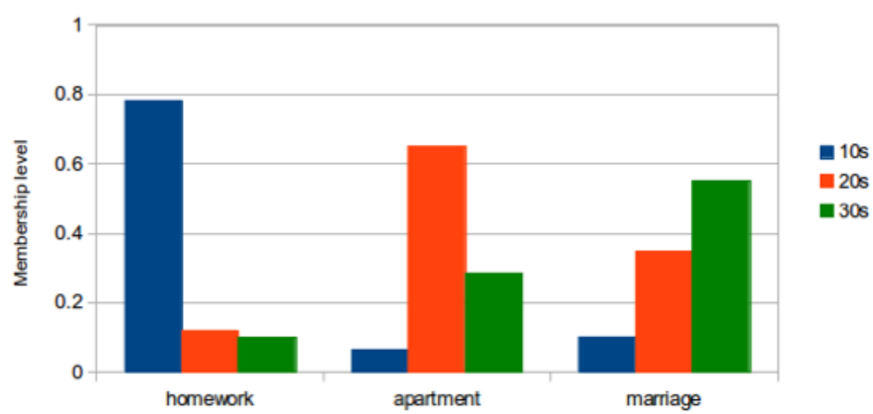
Our proposal: a concise representation

- Terms are represented by their association with profiles
- Document representations are built by combining term representations.

	p_1	.	.	.	p_i
t_1	$wtp_{11}(p_1, t_1)$	.	.	.	$wtp_{i1}(p_i, t_1)$
.	.	.			
.	.		.		
.	.			.	
t_j	$wtp_{1j}(p_1, t_j)$				$wtp_{ij}(p_i, t_j)$

$$\vec{d}_k = \sum_{t_j \epsilon D_k} \frac{tf_{kj}}{len(d_k)} \times \vec{t}_j$$

	p_1	.	.	.	p_i
d_1	$dp_{11}(p_1, d_1)$	.	.	.	$dp_{i1}(p_i, d_1)$
.	.	.			
.	.		.		
.	.			.	
d_j	$dp_{1j}(p_1, d_j)$				$dp_{ij}(p_i, d_j)$



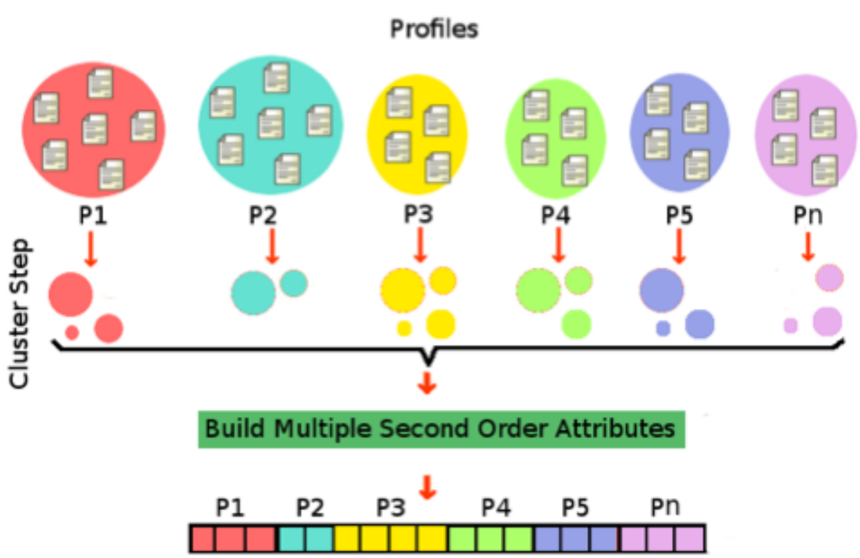
Laboratorio de Tecnologías del Lenguaje
Ciencias Computacionales, INAOE

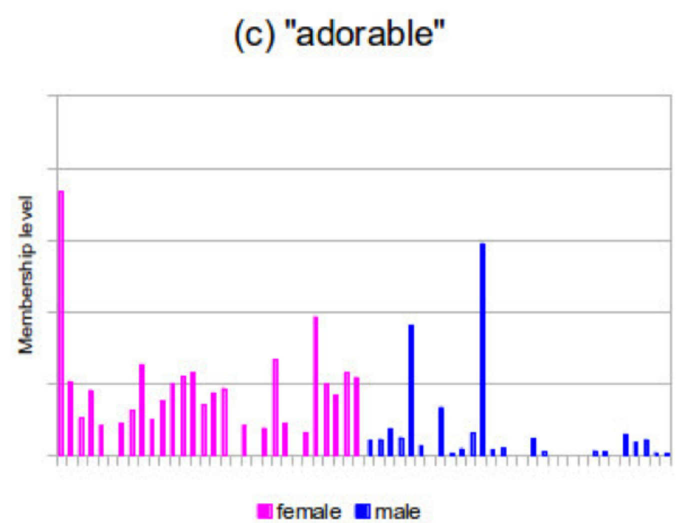
Results at PAN 2013

- We obtained the best results:
  - English: 0.57 (gender), 0.66 (age: 3 ranges)
  - Spanish: 0.63 (gender), 0.66 (age: 3 ranges)
- However, our approach assumes homogeneity among the authors from a given profile. That is not true especially in social media.
- Our solution for PAN 2014: same approach but using information from *subprofiles*.

Subprofile-based representation

- Each profile is clustered in several subprofiles
- Representations for terms and documents are built at subprofile level.
  - That is, there are as many features as subprofiles





Resultados en PAN 2014

- Once again the best performance at PAN.
- The subprofile based representation (n-SOA) was better than the profile-based (SOA) and bag of terms (BoT) representations.
  - In all cases we considered the most frequent 50k terms.

Age and gender prediction in the english dataset.

Dataset	Representation	Blogs Age	Blogs Gender	Twitter Age	Twitter Gender	Social Media Age	Social Media Gender	Reviews Age	Reviews Gender
Train	BoT	45.57	73.87	39.21	71.52	34.30	54.29	31.17	64.87
	1-SOA	46.72	75.44	43.52	70.52	35.81	55.01	32.63	66.75
	n-SOA	**48.07**	**77.96**	**47.97**	**71.98**	**37.00**	**55.36**	**33.92**	**68.05**
Test	n-SOA	39.74	67.95	49.35	72.08	35.52	52.37	33.37	68.09

Age and gender prediction in the English corpus

Dataset	Representation	Blogs Age	Blogs Gender	Twitter Age	Twitter Gender	Social Media Age	Social Media Gender
Train	BoT	43.18	62.50	39.88	62.60	37.65	63.83
	1-SOA	45.33	62.91	41.54	62.01	38.88	64.47
	n-SOA	**48.22**	**63.05**	**43.61**	**62.51**	**41.42**	**65.35**
Test	n-SOA	48.21	58.93	53.33	60.00	45.23	64.84

Two current lines of research

Personality detection

Is it possible to determine the personality of a person by analyzing her social media activity?

Multimodal analysis

Are the images useful for author profiling?
Is their information complementary to the textual data?

POLÍTICA

NOTICIAS DE HOY ESPECIALES MULTIMEDI

USTED ESTÁ AQUÍ: INICIO / POLÍTICA / CREAN MODELO DE IA PARA DET PEDERASTAS EN LAS REDES SOCIALES /

Se produce pornografía infantil en México, según estudio

Crean modelo de IA para detectar pederas las redes sociales

EMIR OLIVARES ALONSO

Periódico La Jornada
Viernes 27 de diciembre de 2013, p. 10

Los delitos cibernéticos contra menores de edad son cada vez más frecuentes en México. Las redes sociales son una de las herramientas más usadas para que pederastas contacten a sus víctimas y puedan obtener desde imágenes de los niños y adolescentes hasta una posible encuentro a solas. El Fondo de

El modelo se probó c de más de 160 mil conv distintos *chats* de redes s que participaron alreded usuarios. Tras el intens datos se logró hallar 2 posibles pederastas.

Un estudio deno *educación a menores en información personal*, ela SEP en 2009, revela que explotación sexual de ni adolescentes a través de I

Forbes MÉXICO INICIO SECCIONES LIST

LO MEJOR Square debe venderse, y rápido Publicado hace 5 horas

Inicio > Tecnología > México desarrolla mecanismo para detectar pederastas en redes sociales

México desarrolla mecanismo para detect pederastas en redes sociales

Acerca de Forbes Staff
Redacción online de la edición mundo. Un equipo de periodist

Escrito por Forbes Staff en noviembre 28, 2013

El proyecto y sus creadores fueron premi

despejado 11/27°

EL UNIVERSAL.mx

PRIMERA

EDICIÓN DI SUSCRÍBASE

Inicio Aviso Oportuno Secciones Suplementos Minuto x Minuto Ed. Impresa Opinión Universal TV C. Deportiva

Nación Metrópoli Edomex Red Política Estados El Mundo Cartera Tu cartera Emprendedor Espectáculos Cultura Estilos

El Universal > Secciones > Primera

Combaten a depredadores de menores

Natalia Gómez Quintero | El Universal
Viernes 06 de diciembre de 2013

Es un programa de computación

ICHA. Investigadores mexicanos desarrollaron un ograma de computación para descubrir a acosadores xuales en internet, cuando se hacen pasar por olescentes para enganchar y abusar de jóvenes. (Foto:)LIGÁN)

MILENIO.COM PUEBLA

Política Firmas Policía Locales Estados Internacional Negocios

CULTURA

INAOE gana concurso en España con proyecto de lingüística forense

El premio otorgado por la Universidad Pompeu Fabra es en memoria de la Dra. M. Teresa Turell i Julia, fundadora y directora de su Laboratorio de Lingüística Forense.

Más de Cultura

Presentarán Los Subterráneos su homenaje a Jack Kerouac

Crean la Primera Banda Sinfónica Femenil Poblana

Manuel Montes, Adrián Pastor y Hugo Jair Escalante, del Laboratorio de Tecnologías del Lenguaje del INAOE. (Foto: Especial)

NOTICIEROS Televisa Noticieros

Deportes Noticieros Televisión Esmas Niños

México DF Estados Mundo Secciones

Coberturas Especiales Por el Planeta Gabriel García Márquez Elige Estar Bien