

Use Fewer Instances of the Letter “i”: Toward Writing Style Anonymization

Andrew W.E. McDonald, Sadia Afroz, Aylin Caliskan,

Ariel Stolerman, and Rachel Greenstadt

PSAL, Drexel University

<http://psal.cs.drexel.edu>

How do you change your
writing style?

Well, Why Change Your Writing Style?

Well, Why Change Your Writing Style?

- Stylometry:
 - Authorship-attribution based on linguistic properties found in text

Well, Why Change Your Writing Style?

- Stylometry:
 - Authorship-attribution based on linguistic properties found in text
 - Stylometric methods can achieve high accuracy across a high number of authors
 - Writeprints: > 90% accuracy for 100 authors (Abbasi and Chen, 2008)
 - Scaling stylometry up to 100,000 authors (Narayanan et al., 2012)

Well, Why Change Your Writing Style?

- Stylometry:
 - Authorship-attribution based on linguistic properties found in text
 - Stylometric methods can achieve high accuracy across a high number of authors
 - Writeprints: > 90% accuracy for 100 authors (Abbasi and Chen, 2008)
 - Scaling stylometry up to 100,000 authors (Narayanan et al., 2012)
- Accuracy and precision of stylometry are increasing

Well, Why Change Your Writing Style?

- Increasingly easier to determine the author of a text

Well, Why Change Your Writing Style?

- Increasingly easier to determine the author of a text
- Harder to express ideas / opinions without taking ownership

Well, Why Change Your Writing Style?

- Increasingly easier to determine the author of a text
- Harder to express ideas / opinions without taking ownership
- Can be a threat to security and privacy

Stylometry: A Threat to Security and Privacy

- Physical
 - Restrictive/oppressive regimes

Stylometry: A Threat to Security and Privacy

- Physical
 - Restrictive/oppressive regimes
- Job
 - Talking against abusive boss

Stylometry: A Threat to Security and Privacy

- Physical
 - Restrictive/oppressive regimes
- Job
 - Talking against abusive boss
- Generally, your writing style could give you away regardless of other precautions taken
 - Tor, VPN, changed MAC address, etc...

Purely Hypothetical? ‡

- Previous examples are purely hypothetical. What about a real example?
- From Inside WikiLeaks by Daniel Domscheit-Berg:
 - “I nudged Julian with my foot. We exchanged glances and started giggling. If someone had run WikiLeaks documents through such a program, he would have discovered that the same two people were behind all the various press releases, document summaries, and correspondence issued by the project. The official number of volunteers we had was also, to put it mildly, grotesquely exaggerated.”



- Okay, so it's necessary to change your writing style to stay anonymous...

- Okay, so it's necessary to change your writing style to stay anonymous...
- Why do we need a tool to do this?

Why Do We Need a Tool for This?

- Machine translation

Why Do We Need a Tool for This?

- Machine translation
 - Either does too little:
 - They passed through the city at noon of the day following.

Why Do We Need a Tool for This?

- Machine translation
 - Either does too little:
 - They passed through the city at noon of the day following.
 - ➔ German ➔ Japanese ➔

Why Do We Need a Tool for This?

- Machine translation
 - Either does too little:
 - They passed through the city at noon of the day following.
 - ➔ German ➔ Japanese ➔
 - They passed the city at noon the following day.

Why Do We Need a Tool for This?

- Machine translation
 - Either does too little:
 - They passed through the city at noon of the day following.
 - ➔ German ➔ Japanese ➔
 - They passed the city at noon the following day.
 - Or does too much:
 - Just remember that the things you put into your head are there forever, he said.

Why Do We Need a Tool for This?

- Machine translation
 - Either does too little:
 - They passed through the city at noon of the day following.
 - ➔ German ➔ Japanese ➔
 - They passed the city at noon the following day.
 - Or does too much:
 - Just remember that the things you put into your head are there forever, he said.
 - ➔ German ➔ Japanese ➔

Why Do We Need a Tool for This?

- Machine translation
 - Either does too little:
 - They passed through the city at noon of the day following.
 - ➔ German ➔ Japanese ➔
 - They passed the city at noon the following day.
 - Or does too much:
 - Just remember that the things you put into your head are there forever, he said.
 - ➔ German ➔ Japanese ➔
 - You are dead, that there always is set, please do not forget what he said.

Why Do We Need a Tool for This?

- ~~Machine translation~~
- Imitation of an author

Why Do We Need a Tool for This?

- ~~Machine translation~~
- Imitation of an author
 - In a small study with 10 participants, not one managed to imitate Cormac McCarthy's writing well enough to fool our Writeprints feature set (not even 10% change in classification)

Why Do We Need a Tool for This?

- ~~Machine translation~~
- Imitation of an author
 - In a small study with 10 participants, not one managed to imitate Cormac McCarthy's writing well enough to fool our Writeprints feature set (not even 10% change in classification)
 - Even if this is managed, it is hard to sustain (e.g. "A Gay Girl in Damascus" – Afroz and Greenstadt, 2012)

Why Do We Need a Tool for This?

- ~~Machine translation~~
- ~~Imitation of an author~~
- Simply obfuscating your writing

Why Do We Need a Tool for This?

- ~~Machine translation~~
- ~~Imitation of an author~~
- Simply obfuscating your writing
 - Brennan and Greenstadt (2009) showed that this can be done while writing the text (not for preexisting writing)

Why Do We Need a Tool for This?

- ~~Machine translation~~
- ~~Imitation of an author~~
- Simply obfuscating your writing
 - Brennan and Greenstadt (2009) showed that this can be done while writing the text (not for preexisting writing)
 - However, the quality of the texts produced was far from scholarly

Why Do We Need a Tool for This?

- ~~Machine translation~~
- ~~Imitation of an author~~
- Simply obfuscating your writing
 - Brennan and Greenstadt (2009) showed that this can be done while writing the text (not for preexisting writing)
 - However, the quality of the texts produced was far from scholarly
 - No way to “know” if you are doing the right thing

Why Do We Need a Tool for This?

- ~~Machine translation~~
- ~~Imitation of an author~~
- ~~Simply obfuscating your writing~~

Anonymouth



- Java based program that uses JStylo and machine learning techniques to attempt to aid users in severing stylometric ties between themselves and a document they authored

The Goals

- An efficient, usable, and effective tool to allow people to express their thoughts anonymously

The Goals

- An efficient, usable, and effective tool to allow people to express their thoughts anonymously
- Make clear that stylometry can be fooled, and therefore it cannot be relied upon absolutely

The Goals

- An efficient, usable, and effective tool to allow people to express their thoughts anonymously
- Make clear that stylometry can be fooled, and therefore it cannot be relied upon absolutely
- A tool that provides a usable interface between the outcomes from machine learning analytics and a user

Key Contributions

- JStylo-Anonymouth framework
 - Authorship attribution
 - Identifies changes required for document anonymization relative to a corpus
 - Assists the user making necessary changes accordingly

Key Contributions

- JStylo-Anonymouth framework
 - Authorship attribution
 - Identifies changes required for document anonymization relative to a corpus
 - Assists the user making necessary changes accordingly
- User study showed that Anonymouth is effective in determining what needs to be changed to achieve anonymization
 - But has trouble telling the user how to make the changes

Problem Statement

- Author A has a document D to anonymize

Problem Statement

- Author A has a document D to anonymize
 - P : set of labeled documents written by A

Problem Statement

- Author A has a document D to anonymize
 - P : set of labeled documents written by A
 - O : set of labeled documents written by N other authors (“blend-in” corpus)

Problem Statement

- Author A has a document D to anonymize
 - P : set of labeled documents written by A
 - O : set of labeled documents written by N other authors (“blend-in” corpus)
 - F : set of linguistic features to extract

Problem Statement

- Author A has a document D to anonymize
 - P : set of labeled documents written by A
 - O : set of labeled documents written by N other authors (“blend-in” corpus)
 - F : set of linguistic features to extract
 - C : classifier trained on $P \cup O$

Problem Statement

- Author A has a document D to anonymize
 - P : set of labeled documents written by A
 - O : set of labeled documents written by N other authors (“blend-in” corpus)
 - F : set of linguistic features to extract
 - C : classifier trained on $P \cup O$
- Goal: create D' from D where the extracted features from F are sufficiently changed such that:

$$\Pr[C(D') = A] \leq \frac{1}{N+1}$$

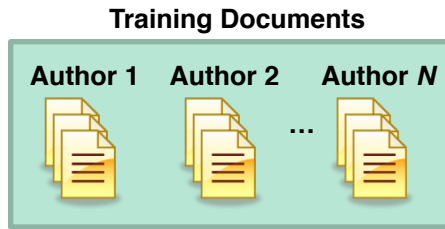
Framework

- JStylo
 - authorship attribution
- Anonymouth
 - authorship anonymization

JStylo

- Standalone platform for authorship attribution
- Underlying feature extraction and authorship attribution engine of Anonymouth
- NLP for feature extraction
- Supervised machine learning:
 - Trains on documents of known candidate authors
 - Tests anonymous documents to attribute authorship
- Phases:
 - Problem set definition
 - Feature extraction
 - Classifiers selection
 - Analysis
- Powered by JGAAP, Weka

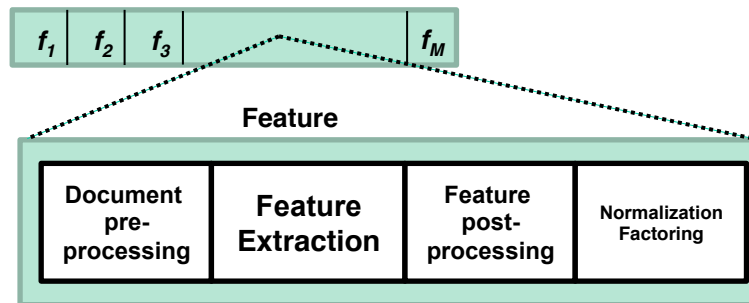
1. Problem Definition



Test Documents



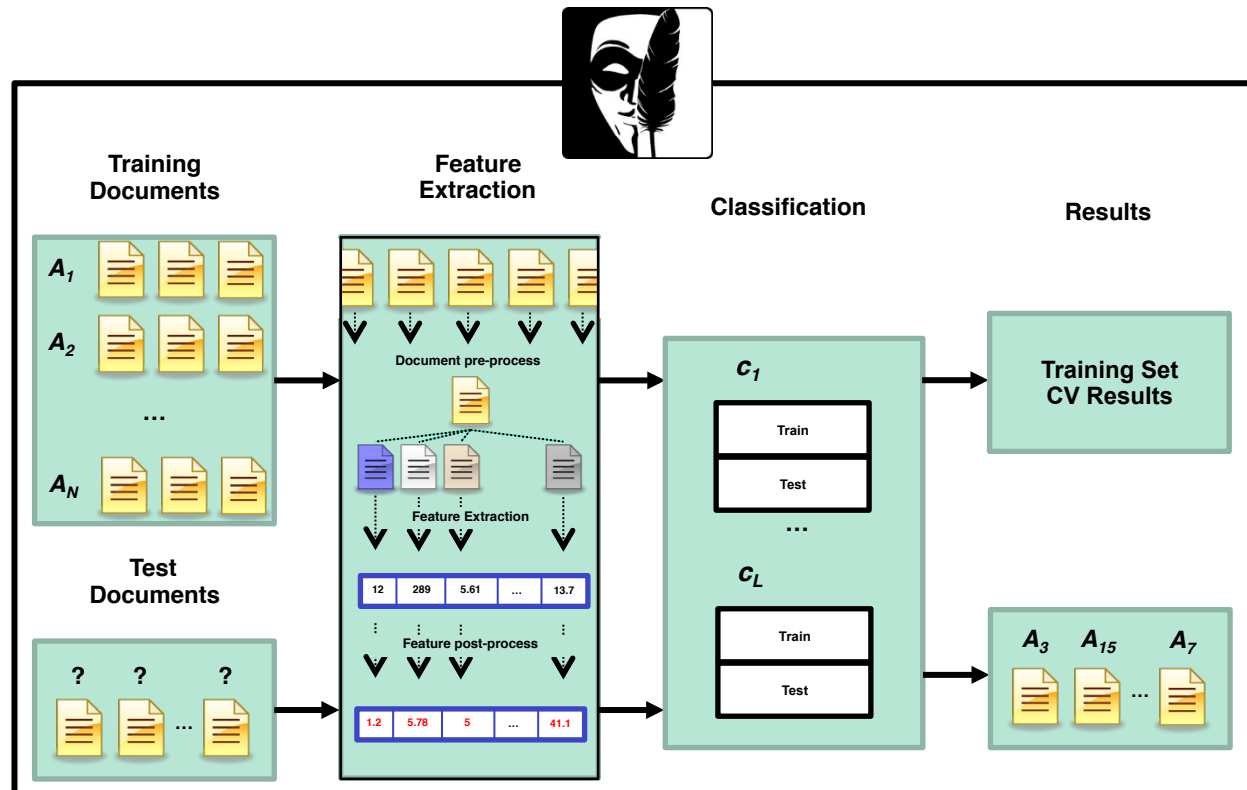
2. Feature Selection



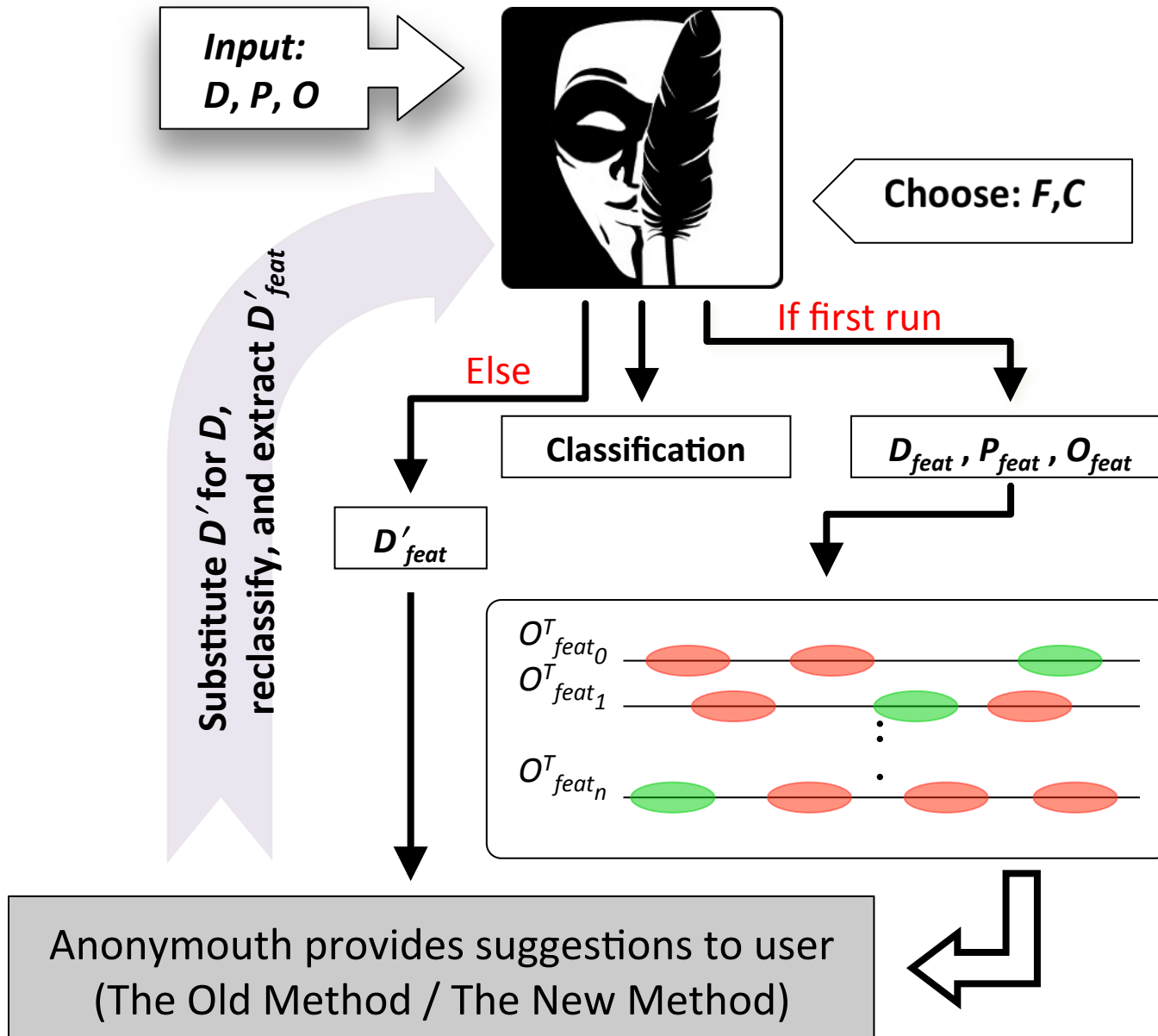
3. Classifiers Selection



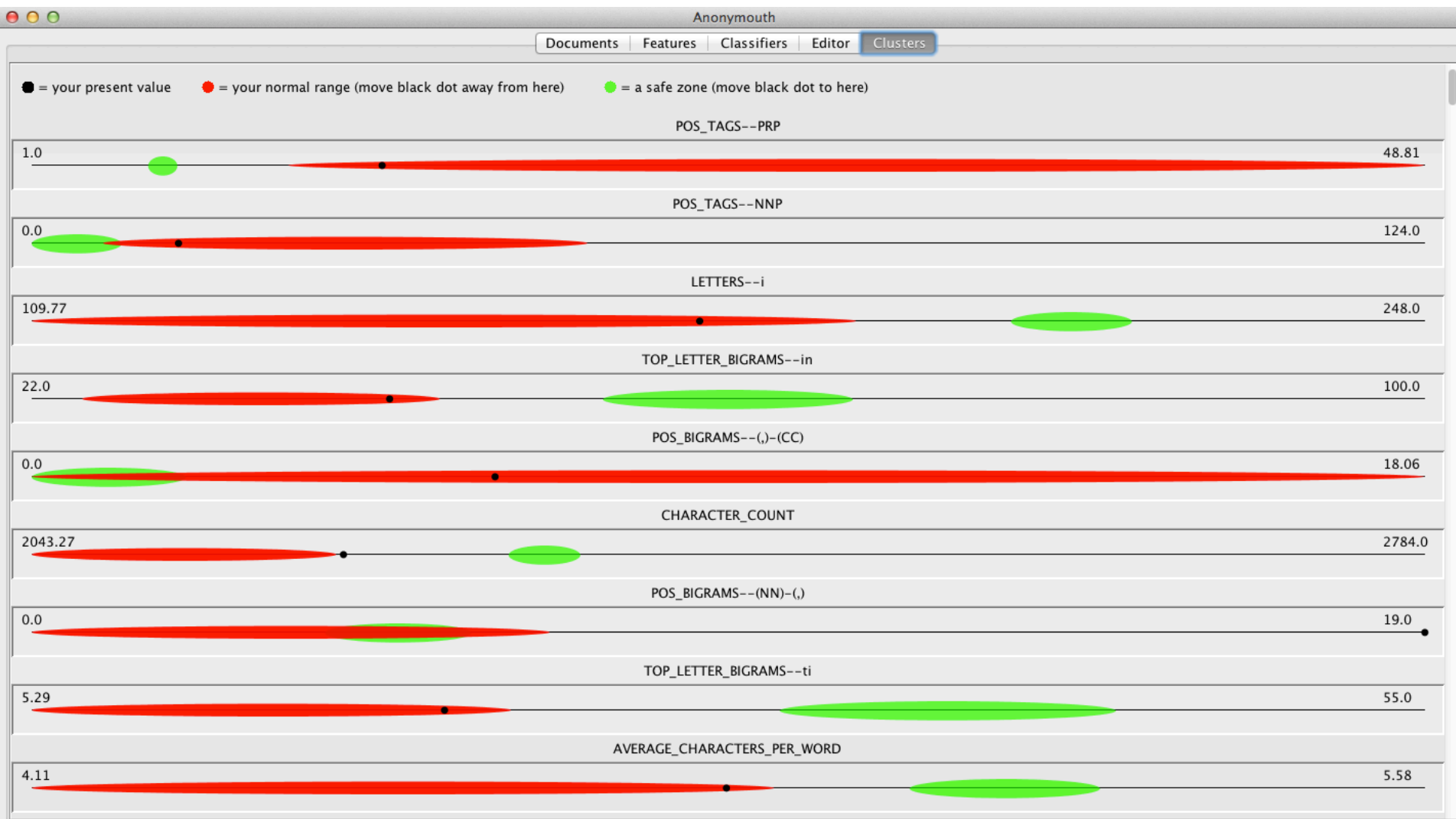
4. Analysis



General Design of Anonymouth



Anonymouth's Clusters (updated display)

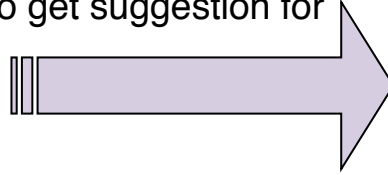


The green circles comprise the target cluster group, T . The red circles are the author's confidence interval for each feature, and the black dot on each plot is the present value of that feature in the author's document to anonymize

First Try: The Old Method

O_{feat_0}
O_{feat_1}
\vdots
O_{feat_n}

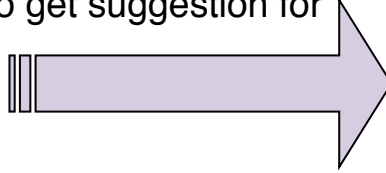
User selects feature
to get suggestion for



First Try: The Old Method

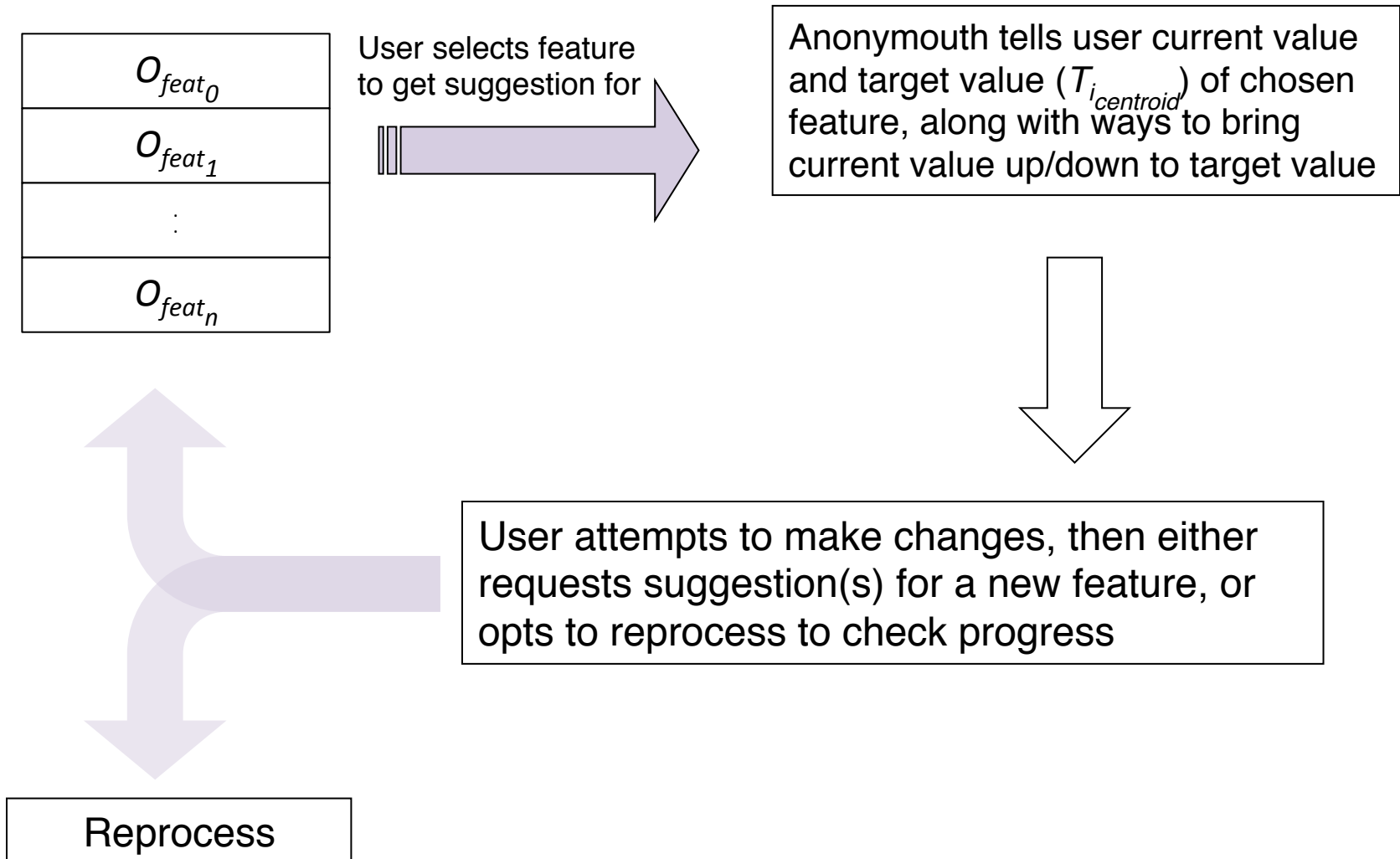
O_{feat_0}
O_{feat_1}
\vdots
O_{feat_n}

User selects feature
to get suggestion for



Anonymouth tells user current value
and target value ($T_{i_{centroid}}$) of chosen
feature, along with ways to bring
current value up/down to target value

First Try: The Old Method



Results of User Study Using The Old Method

- When people are able to do what is asked of them, they are able to anonymize their writing

Results of User Study Using The Old Method

- When people are able to do what is asked of them, they are able to anonymize their writing
- 8/10 participants were able to anonymize their writing with respect to JStylo's Basic-9 feature set

Results of User Study Using The Old Method

- When people are able to do what is asked of them, they are able to anonymize their writing
- 8/10 participants were able to anonymize their writing with respect to JStylo's Basic-9 feature set
- 7/10 were able to achieve a classification less than or equal to random chance

Results of User Study Using The Old Method

- When people are able to do what is asked of them, they are able to anonymize their writing
- 8/10 participants were able to anonymize their writing with respect to JStylo's Basic-9 feature set
- 7/10 were able to achieve a classification less than or equal to random chance
- But, no users were able to anonymize their writing using JStylo's Writeprints feature set

Results of User Study Using The Old Method

- However, changing $\approx 14\%$ (≈ 100) of the features Anonymouth indicated, 8/10 user's documents were classified as having been written by a different author with 95% probability (Jstylo's Writeprints feature set)

Results of User Study Using The Old Method

- However, changing $\approx 14\%$ (≈ 100) of the features Anonymouth indicated, 8/10 user's documents were classified as having been written by a different author with 95% probability (Jstylo's Writeprints feature set)
- Revealed that initial approach needed revision

Results of User Study Using The Old Method

- However, changing $\approx 14\%$ (≈ 100) of the features Anonymouth indicated, 8/10 user's documents were classified as having been written by a different author with 95% probability (Jstylo's Writeprints feature set)
- Revealed that initial approach needed revision
- Conclusion: Anonymouth was unusable, but core methodology sound
- Also, asking users to select cluster groups is problematic

Editor – Basic-9 Feature Set

DocumentsFeaturesClassifiersEditorClusters

OriginalOriginal->1

Along with the coffee the waitress kindly pours me a shot of a golden spirit which she informs me is a specialty from somewhere in the North. As I am never one to offend with the refusal of such hospitality, I graciously oblige. The spirit tasted subtly of anise, almost like liquid fennel. It is wonderful. My only regret is that I cannot remember the name.

From lunch I take the metro directly to the Atocha station to see the Reina Sofia museum; I'm told it has free admission on Saturday afternoons. I had visited the museum on my previous visit to Madrid, but it really is one of my favorites. Apparently I am not the only one with the bright idea of saving a few Euros on admission; the queue is huge. It ends up being a 30 minute wait, in the rain, under my flimsy little hotel loaner umbrella. It is worth it, though.

After a few hours ogling the likes of Dalí and Picasso, I zip over to Sol to try and find a famed sherry bar called La Venencia. Unfortunately the bar hasn't yet opened, so I walk another few blocks with the intention of completing my tapas crawl from two days prior. The tapas places haven't yet reopened after siesta, though, so I am out of luck. The single eatery that is open is the Museo del Jamón. Despite its name, it's really more of a franchised bar/restaurant than a museum. It's a very touristy, but I am hungry for a snack so I enter and order a plate of Iberica. The Japanese businessman next to me is just finishing a plate of his own and wishes to ask for the bill. He whips out his Madrid tour guide, studies it for what seems like five minutes.

The train from Madrid to Lisbon departs at 22:45. It's an eleven hour journey. Due to extenuating circumstances including—but not limited to—the perihelion of Mercury, I have somehow scored a first class cabin all to myself, including a meal in the dining car. I settle into my tiny cabin; the only "first class" aspect of which is that it has an ensuite bathroom, which was not entirely unexpected.

I am the only unaccompanied person in the first class car, and seemingly the only one below retirement age. The other three cabins are all likewise occupied by non-Europeans. My neighbors, I learn, are a couple that live in Jamaica. The husband, one Clinton P. Chin, J.P. (which I assume stands for "Justice of the Peace") is the chairman of the Chinese Twinning Commission for Hangzhou ? Montego Bay and Zhejiang Province. Apparently, the Twinning Commission oversees relations and commerce between the aforementioned regional pairs. I do not discover this until His Worship, the Honourable Mr. Chin gave me his business card, so I do not have a chance to ask him what types of commerce occurs between Hangzhou and Zhejiang and their respective Rasta relations.

Results of **Last** Document's Classification (% probability of authorship per author)

g	s	p	m	~* you *~	k	h
20.04	0.01	0.0	0.49	79.45	0.0	0.0
Actual Author:	~* you *~					

Unfortunately, your document seems to have been written by: ~* you *~

Highlight: UNIQUE_WORDSspecific value

User Editing... Waiting to "Re-process"

Re-processClear HighlightsDictionaryVerboseSave...Exit

Suggestion

You should try to decrease your unique word count to 252.88 (from its present value of '303.00') by replacing some single use words with less than 3 syllables with words that have already been used and have 3 or more syllables.

UNIQUE_WORDS_COUNT :

Present Value: 303.0

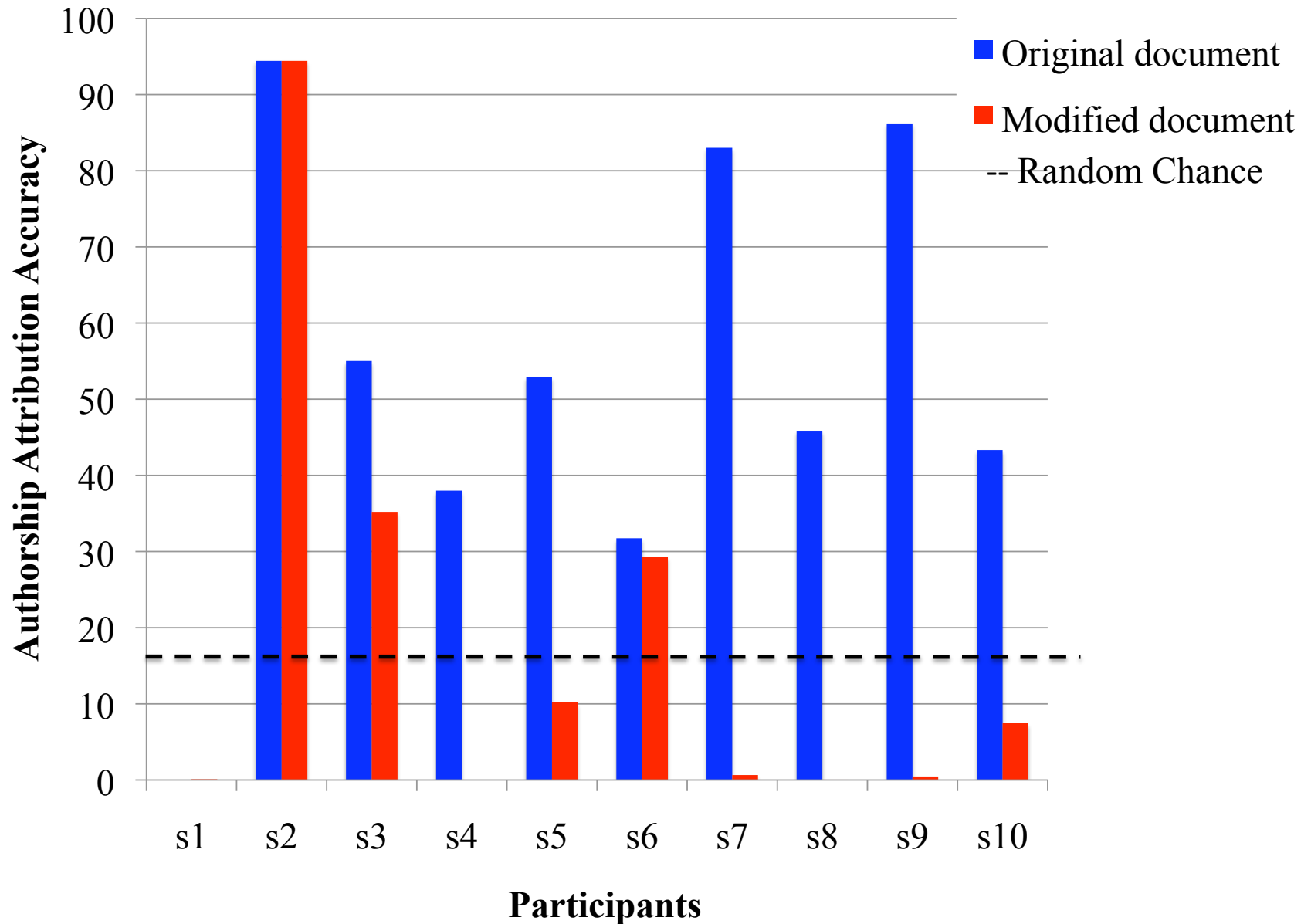
Target Value: 252.875

List of Suggestions

No.	Feature Name
1	SENTENCE_COUNT
2	AVERAGE_SYLLABLES_IN_WORD
3	CHARACTER_SPACE
4	LETTER_SPACE
5	UNIQUE_WORDS_COUNT

Basic-9 Feature Set Results

Anonymization in terms of original background corpus



Editor – Writeprints Feature Set

DocumentsFeaturesClassifiersEditorClusters

OriginalOriginal->1

Along with the coffee the waitress kindly pours me a shot of a golden spirit which she informs me is a specialty from somewhere in the North. As I am never one to offend with the refusal of such hospitality, I graciously oblige. The spirit tasted subtly of anise, almost like liquid fennel. It is wonderful. My only regret is that I cannot remember the name.

From lunch I take the metro directly to the Atocha station to see the Reina Sofia museum; I'm told it has free admission on Saturday afternoons. I had visited the museum on my previous visit to Madrid, but it really is one of my favorites. Apparently I am not the only one with the bright idea of saving a few Euros on admission: the queue is huge. It ends up being a 30 minute wait, in the rain, under my flimsy little hotel loaner umbrella. It is worth it, though.

After a few hours ogling the likes of Dalí and Picasso, I zip over to Sol to try and find a famed sherry bar called La Venencia. Unfortunately the bar hasn't yet opened, so I walk another few blocks with the intention of completing my tapas crawl from two days prior. The tapas places haven't yet reopened after siesta, though, so I am out of luck. The single eatery that is open is the Museo del Jamón. Despite its name, it's really more of a franchised bar/restaurant than a museum. It's a very touristy, but I am hungry for a snack so I enter and order a plate of Iberica. The Japanese businessman next to me is just finishing a plate of his own and wishes to ask for the bill. He whips out his Madrid tour guide, studies it for what seems like five minutes.

The train from Madrid to Lisbon departs at 22:45. It's an eleven hour journey. Due to extenuating circumstances including—but not limited to—the perihelion of Mercury, I have somehow scored a first class cabin all to myself, including a meal in the dining car. I settle into my tiny cabin; the only "first class" aspect of which is that it has an ensuite bathroom, which was not entirely unexpected.

I am the only unaccompanied person in the first class car, and seemingly the only one below retirement age. The other three cabins are all likewise occupied by non-Europeans. My neighbors, I learn, are a couple that live in Jamaica. The husband, one Clinton P. Chin, J.P. (which I assume stands for "Justice of the Peace") is the chairman of the Chinese Twinning Commission for Hangzhou ? Montego Bay and Zhejiang Province. Apparently, the Twinning Commission oversees relations and commerce between the aforementioned regional pairs. I do not discover this until His Worship, the Honourable Mr. Chin gave me his business card, so I do not have a chance to ask him what types of commerce occurs between Hangzhou and Zhejiang and their respective Rasta relations.

Results of **Last** Document's Classification (% probability of authorship per author)

g	s	p	m	~* you *~	k	h
0.57	0.07	0.01	0.01	99.33	0.01	0.0
Actual Author:	~* you *~					

Unfortunately, your document seems to have been written by: ~* you *~

Highlight: None specific value

User Editing... Waiting to "Re-process"

Re-processClear HighlightsDictionaryVerboseSave...Exit

Suggestion

The description of the part of speech tag 'PRP' is: Personal pronoun, and the description of 'VBP' is: Verb, non-3rd person singular present. Currently, you have too few of these part of speech bigrams (pairs) in your document. You have '16.0', while you should have '24.2'. The highlighted words are examples of word pairs that would be/ are tagged as (PRP)-(VBP) in the context that they are used. Try to use these words in their context as a guideline to add 8.2 word pairs that would be tagged as '(PRP)-(VBP)'.

POS_BIGRAMS (PRP)-(VBP):

Present Value: 16.0

Target Value: 24.2

List of Suggestions

No.	Feature Name
1	LETTERS i
2	AVERAGE_CHARACTERS_PER_WORD
3	UPPERCASE_LETTERS_PERCENTAGE
4	FUNCTION_WORDS i
5	WORDS i
6	LETTERS c
7	POS_BIGRAMS (PRP)-(VBP)
8	POS_TAGS NNP
9	LETTERS n
10	POS_TAGS PRP\$
11	CHARACTER_COUNT
12	TOP_LETTER_BIGRAMS io
13	WORD_LENGTHS 1
14	LETTERS PERCENTAGE

Next Steps: The New Method

- Need to present user with more feasible task
- Tweak / change cluster ordering algorithms to eliminate the need for user to select a cluster group
- Present document sentence by sentence, rather than all at once
- Ask to change words rather than specific features
- Allow user to shuffle and remove words to gain a different perspective / help user to see each sentence from a different perspective
- Find “best” synonyms for words
- (Possibly) analyze point of view (1st, 2nd, 3rd person), tense, and verb conjugation habits

Extract POS tags using Stanford POS Tagger, wrap each “word” in an object which keeps track of how many features from O_{feat} were found in that word.

Reprocess

User requests next sentence or opts to reprocess

For each word compute:

$$\text{Anonymity Index} = \sum_{i=0}^n \left(\frac{a_i}{T} \right) \times (g_i) \times (p_i)$$

Where: a_i = appearances of feature i in word; T = total features found in word; g_i = Information Gain of feature i ; and p_i = percent change needed of feature i

Negative
Anonymity Index

Positive
Anonymity Index

Words to remove

Words to add

Present sentence to user with lists of words to add and to remove

User may choose to “shuffle” sentence, with or without stripping “words to remove”, which randomly reorganizes remaining words in sentence edit box

When user is satisfied with changes to current sentence

Going Forward

- Clustering is unreliable
- Usability:
 - What information should be given to the user?
 - What is the best way to get the user to make the changes needed to the document?
 - Testing our new method

Questions?

- Contact information:
 - Andrew McDonald
 - awm32@cs.drexel.edu
 - Sadia Afroz
 - sa499@cs.drexel.edu
 - Aylin Caliskan
 - ac993@cs.drexel.edu
 - Ariel Stolerman
 - ams573@cs.drexel.edu
 - Rachel Greenstadt
 - greenie@cs.drexel.edu
- Drexel PSAL website (to download JStylo and Anonymouth):
 - <https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth>

How Does it Work?

- User inputs three sets of documents
 - \mathbf{D} , \mathbf{P} , and \mathbf{O}
- He chooses
 - A feature set, \mathbf{F} ,
 - A classifier, \mathbf{C}
- JStylo extracts features from \mathbf{D} , \mathbf{P} , and \mathbf{O} , creating \mathbf{D}_{feat} , \mathbf{P}_{feat} , and \mathbf{O}_{feat} , respectively, and delivers a classification of document \mathbf{D} among $\mathbf{P} \cup \mathbf{O}$
- Anonymouth individually clusters the values of a single feature for all documents represented by \mathbf{O}_{feat}^T

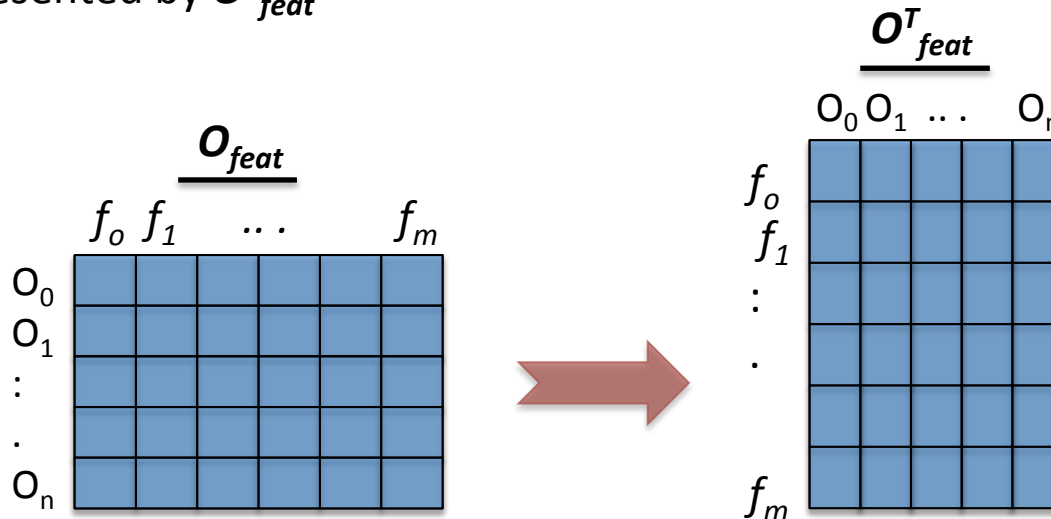
How Does it Work?

- User inputs three sets of documents
 - \mathbf{D} , \mathbf{P} , and \mathbf{O}
- He chooses
 - A feature set, \mathbf{F} ,
 - A classifier, \mathbf{C}
- JStylo extracts features from \mathbf{D} , \mathbf{P} , and \mathbf{O} , creating \mathbf{D}_{feat} , \mathbf{P}_{feat} , and \mathbf{O}_{feat} , respectively, and delivers a classification of document \mathbf{D} among $\mathbf{P} \cup \mathbf{O}$
- Anonymouth individually clusters the values of a single feature for all documents represented by \mathbf{O}_{feat}^T

	<u>\mathbf{O}_{feat}</u>				
	f_0	f_1	\dots		f_m
O_0					
O_1					
$:$					
\cdot					
O_n					

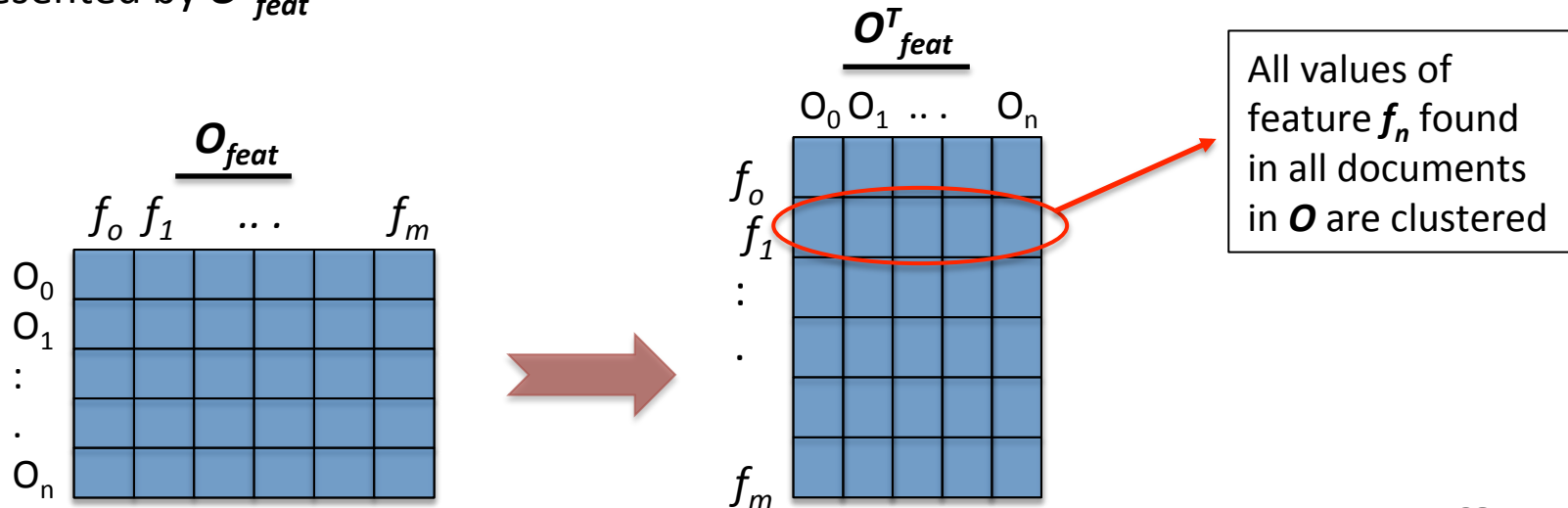
How Does it Work?

- User inputs three sets of documents
 - \mathbf{D} , \mathbf{P} , and \mathbf{O}
- He chooses
 - A feature set, \mathbf{F} ,
 - A classifier, \mathbf{C}
- JStylo extracts features from \mathbf{D} , \mathbf{P} , and \mathbf{O} , creating \mathbf{D}_{feat} , \mathbf{P}_{feat} , and \mathbf{O}_{feat} , respectively, and delivers a classification of document \mathbf{D} among $\mathbf{P} \cup \mathbf{O}$
- Anonymouth individually clusters the values of a single feature for all documents represented by \mathbf{O}_{feat}^T



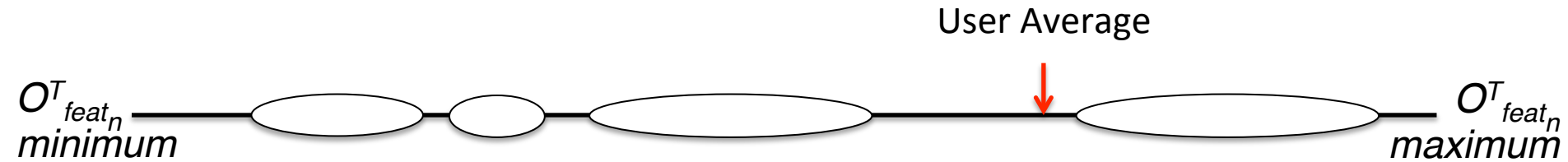
How Does it Work?

- User inputs three sets of documents
 - D , P , and O
- He chooses
 - A feature set, F ,
 - A classifier, C
- JStylo extracts features from D , P , and O , creating D_{feat} , P_{feat} , and O_{feat} , respectively, and delivers a classification of document D among $P \cup O$
- Anonymouth individually clusters the values of a single feature for all documents represented by O_{feat}^T

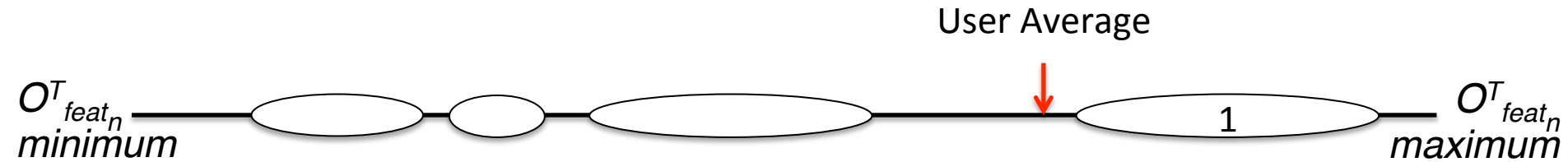


- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)

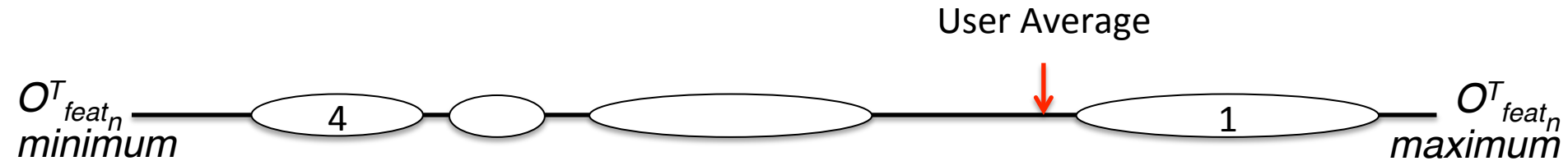
- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)



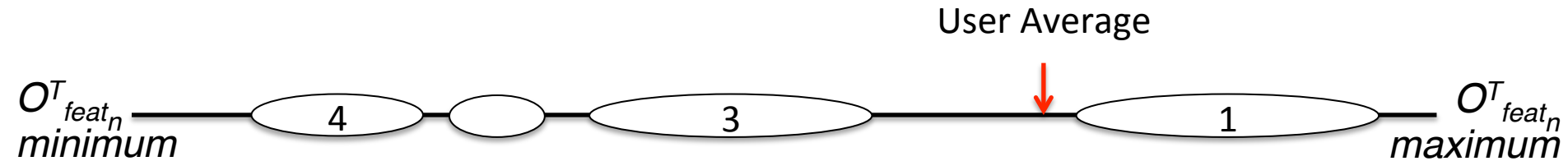
- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)



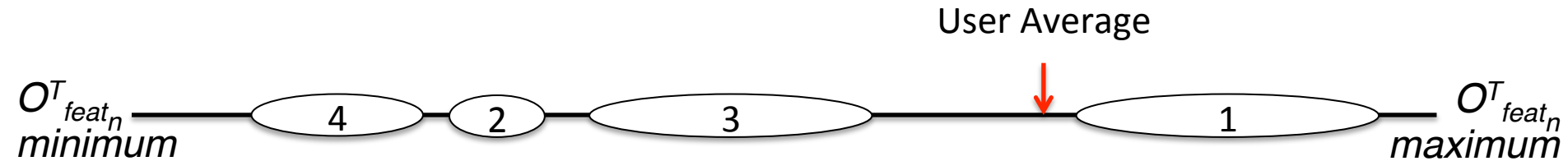
- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)



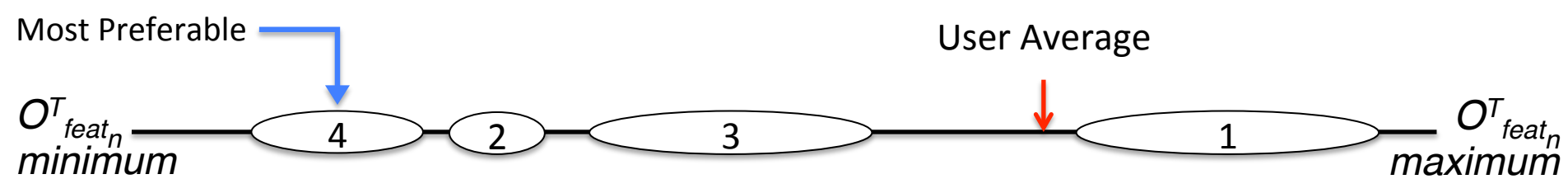
- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)



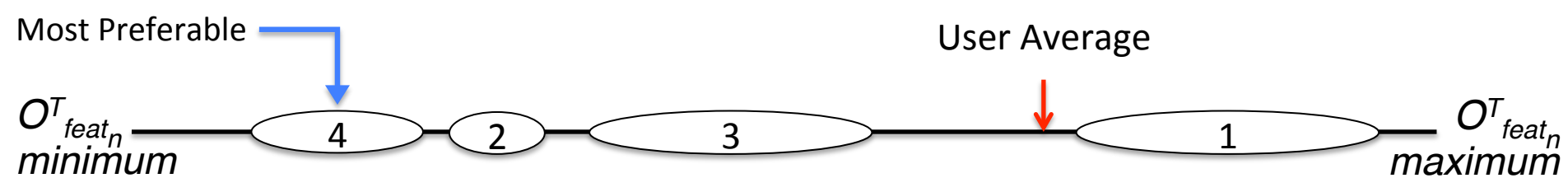
- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)



- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)




- Primary preference calculation orders each feature's clusters based upon number of elements and distance from user's average (for that feature)



- Once all features have been clustered and ordered, cluster groups are assembled
 - Each cluster group represents at least one document

- Cluster groups are ordered via a secondary preference calculation which gives the greatest preference to the cluster group who's clusters for the highest ranking features in terms of information gain are the most preferable clusters for their respective features

Greatest
Information Gain



	f_0	f_1	f_2	..	.	f_n
CG_0	5	3	6	..	.	2
CG_1	5	3	5	..	.	5
CG_2	4	2	4	..	.	4
:	:	:	:	..	.	:
.
CG_z	1	1	1	..	.	1


- Cluster groups are ordered via a secondary preference calculation which gives the greatest preference to the cluster group who's clusters for the highest ranking features in terms of information gain are the most preferable clusters for their respective features


Greatest
Information Gain

Target cluster group, T →

	f_0	f_1	f_2	..	.	f_n
CG_0	5	3	6	..	.	2
CG_1	5	3	5	..	.	5
CG_2	4	2	4	..	.	4
:	:	:	:	..	.	:
.
CG_z	1	1	1	..	.	1

- Cluster groups are ordered via a secondary preference calculation which gives the greatest preference to the cluster group who's clusters for the highest ranking features in terms of information gain are the most preferable clusters for their respective features

		Greatest Information Gain			Least Information Gain		
		f_0	f_1	f_2	..	.	f_n
Target cluster group, T 	CG_0	5	3	6	..	.	2
	CG_1	5	3	5	..	.	5
	CG_2	4	2	4	..	.	4
	:	:	:	:	..	.	:

Least preferred cluster group 	CG_z	1	1	1	..	.	1

- Cluster groups are ordered via a secondary preference calculation which gives the greatest preference to the cluster group who's clusters for the highest ranking features in terms of information gain are the most preferable clusters for their respective features

Greatest
Information Gain

Target cluster group, T →

	f_0	f_1	f_2	..	.	f_n
CG_0	5	3	6	..	.	2
CG_1	5	3	5	..	.	5
CG_2	4	2	4	..	.	4
:	:	:	:	..	.	:
.
CG_z	1	1	1	..	.	1

Each cluster group's clusters must each contain one feature from at least a single document in \mathcal{O}

- Cluster groups are ordered via a secondary preference calculation which gives the greatest preference to the cluster group who's clusters for the highest ranking features in terms of information gain are the most preferable clusters for their respective features

Greatest Information Gain

Target cluster group, T

T contains the most preferable clusters from the most important features

	f_0	f_1	f_2	..	.	f_n
CG_0	5	3	6	..	.	2
CG_1	5	3	5	..	.	5
CG_2	4	2	4	..	.	4
:	:	:	:	..	.	:
.
CG_z	1	1	1	..	.	1

- Cluster groups are ordered via a secondary preference calculation which gives the greatest preference to the cluster group who's clusters for the highest ranking features in terms of information gain are the most preferable clusters for their respective features

Features with lower Information Gain won't effect the document as much, so they don't effect the secondary preference calculation as much either

Greatest Information Gain

Target cluster group, T

	f_0	f_1	f_2	..	.	f_n
CG_0	5	3	6	..	.	2
CG_1	5	3	5	..	.	5
CG_2	4	2	4	..	.	4
:	:	:	:	..	.	:
.
CG_z	1	1	1	..	.	1