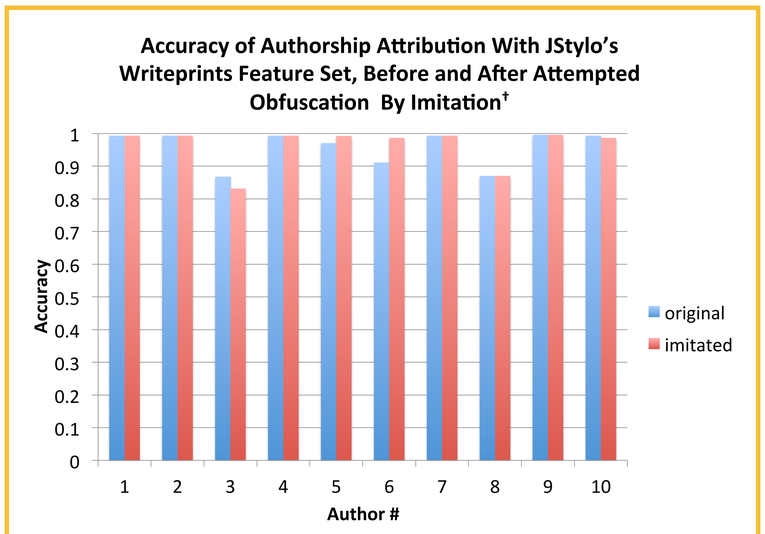
Anonymouth: Toward Authorship Anonymization

Andrew W.E. McDonald, Hoa Vu, Marc Barrowclift, Jeff Ulman, and Rachel Greenstadt

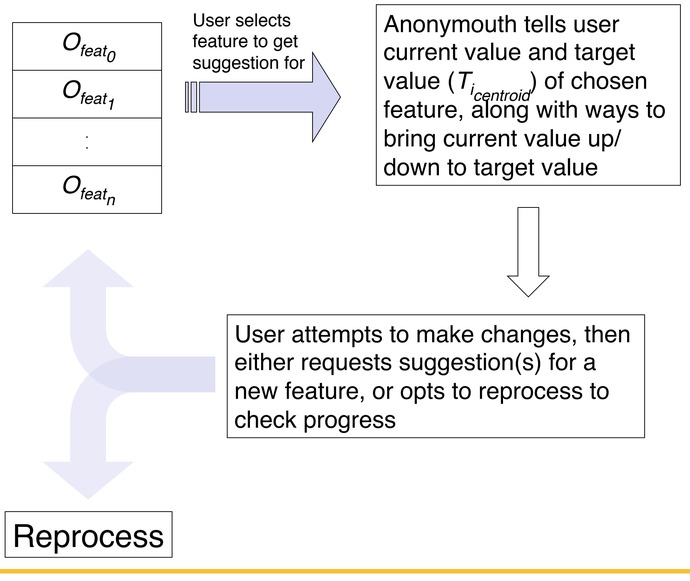
Motivation

With the growing popularity of using stylometric techniques to determine authorship of typed text, it is important to allow people the option of separating themselves from their written work so that they may express their ideas without fear of repercussions from doing so. While it is possible to change one's writing style with no aid, it seems not to be a trivial task to do so while retaining the original affect of the piece. Furthermore, it must be made clear that one cannot rely absolutely on stylometry to determine the author of a work, because as Anonymouth shows, it is possible to alter the features of one's writing; thereby masking the true author of a text.

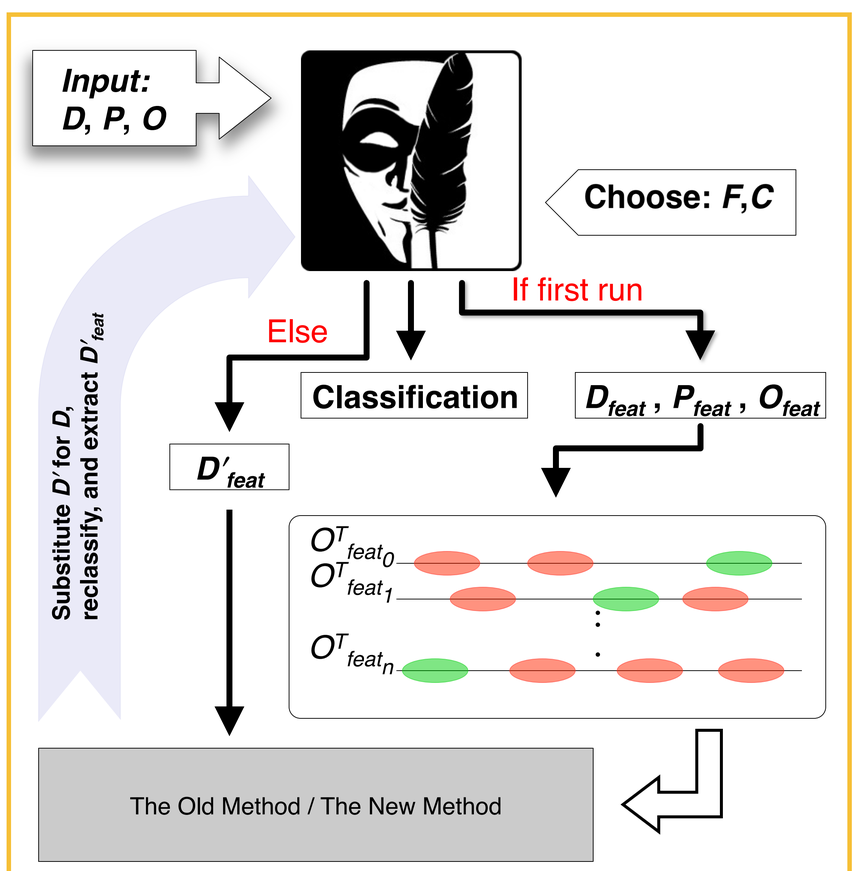


†Experiment performed by, and results obtained from, Sadia Afroz, Drexel University

The Old Method



Overview of Program Flow



How Anonymouth Works

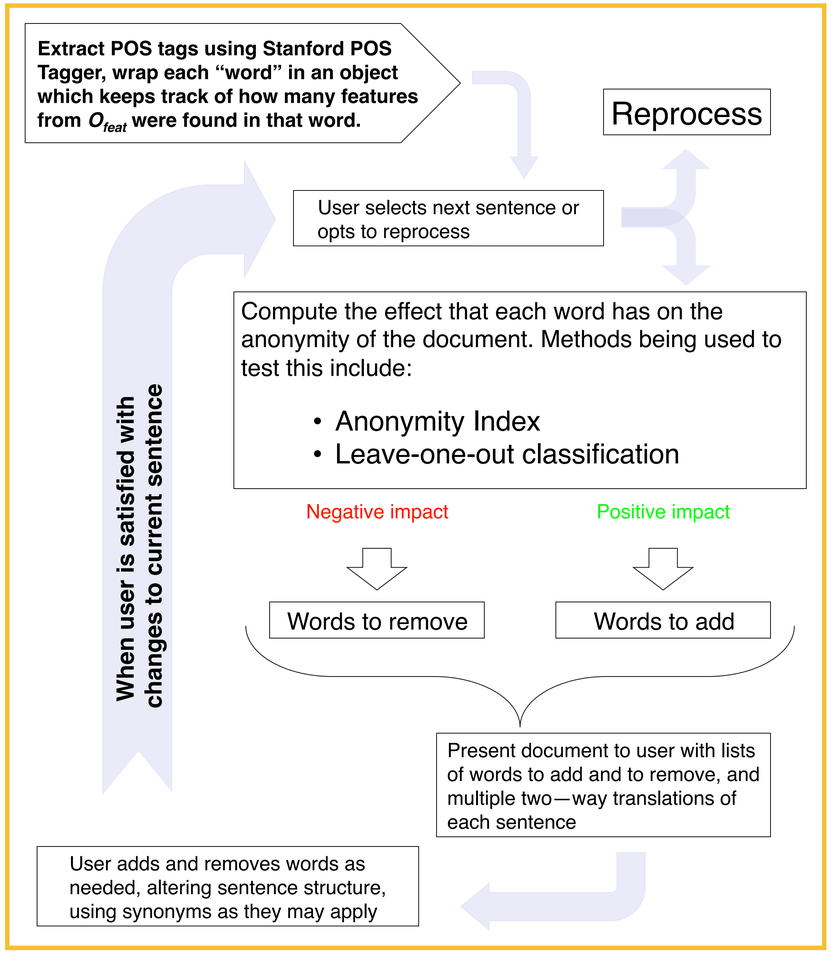
The user, *A*, inputs three sets of documents: a singleton, *D* , to anonymize; his preexisting writing, *P*, and a set *O* of *N* other authors. *A* chooses a feature set, *F*, and classifier, *C*. JStylo extracts features from *D*, *P*, and *O*, respectively creating, D_{feat} , P_{feat} , and O_{feat} . It then classifies *D* among {*P* ∪ *O*}. Anonymouth clusters each feature, $O^T_{feat_i}$, from all documents in *O,* and assigns *D* a target cluster group, *T*, such that at least one document in *O* is represented by the clusters in *T*; preferably with each cluster falling outside of *A*'s confidence interval for each respective feature, giving the greatest importance to those features with the highest information gain. Cluster group selection is accomplished via primary and secondary cluster group preference calculations. The flow then divides into The Old Method‡, and The New Method, however both approaches aim to aid the user, *A*, to create a document, *D′*, that JStylo will classify as having been written by *A* with probability less than or equal to random chance, using D'_{feat} , P_{feat} , and O_{feat} .

‡The Old Method also excluded further consideration of elements of O_{feat} such that $O_{feat_i} \notin \{P_{feat} \cup O_{feat}\}$, and required the user, *A*, to select his own cluster group.

An Authorship Anonymization Platform

Anonymouth is a Java based program that aims to aid a user in stylometrically anonymizing his written text with respect to a chosen feature set, classifier, and reference corpora via machine learning techniques. It uses the output classification from JStylo to determine a baseline authorship classification, and then proceeds to aid the user in changing his document's classification. The aid to the user came in the form of suggestions regarding how to change specific features; though after a user study suggesting no one could effectively implement the suggestions given (and change their classification), it seemed a new approach was in order.

The New Method (Developing)



References

[1] McDonald, A. W. E., Afroz, S., Caliskan, A., Stolerman, A. and Greenstadt, R.: *Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization* (2012)

[2] Brennan, M. and Greenstadt, R.: *Practical Attacks Against Authorship Recognition Techniques* (2009)