***Anonymize v1.1 alpha***

Author: Ralf Hauenschild – ralf\_hauenschild@gmx.de

**GitHub-Repository:**

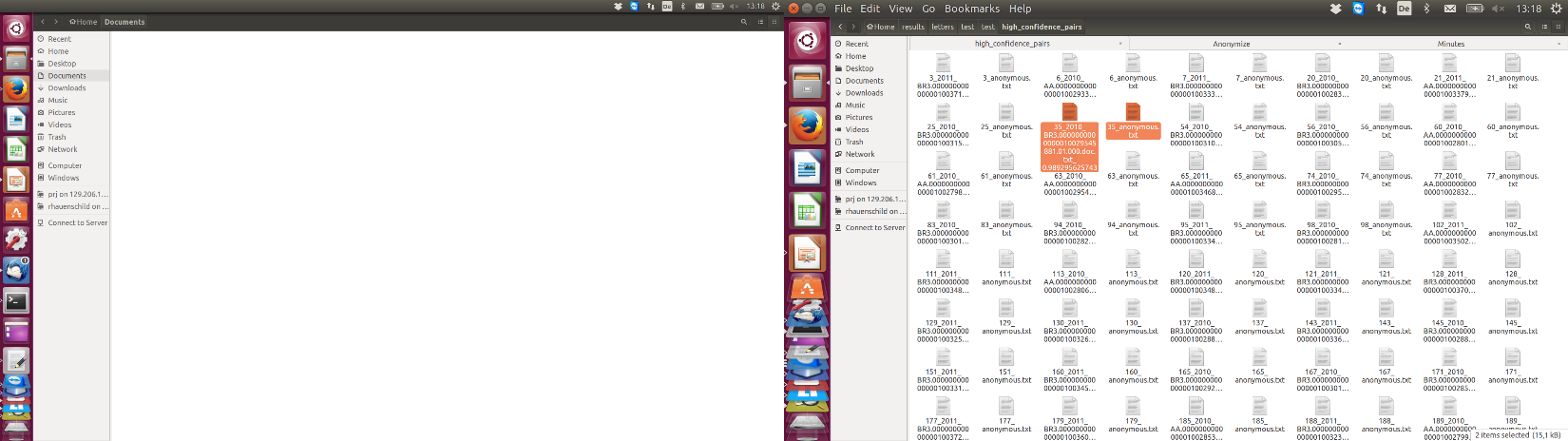
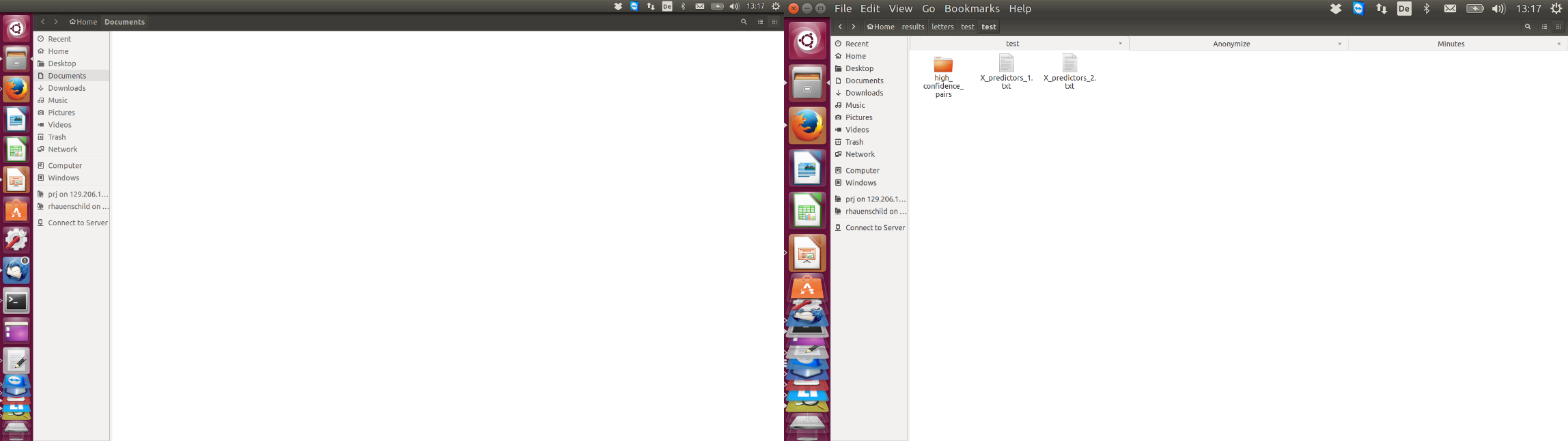
git clone <https://User@github.com:/dieterich-lab/Anonymize>

**Dependencies:** *Python 2.7*, *numpy, matplotlib*

**Usage:**

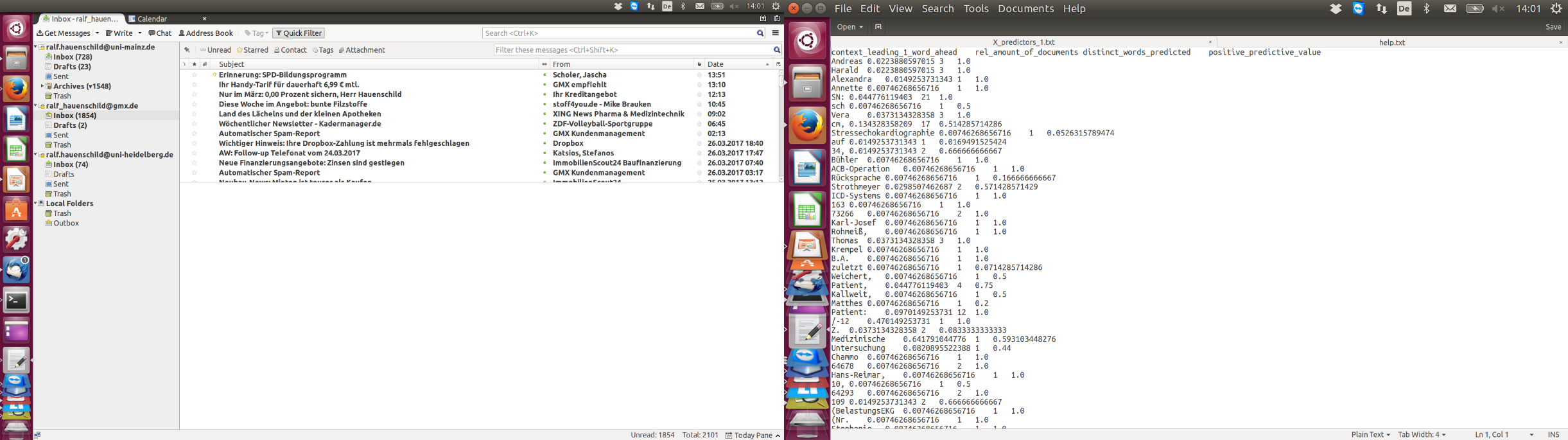
1. **Preparation:** Extract words leading X-ed out content and store them as potential predictors of sensitive content in unseen data. Also, store X-ed out words along with their X-out frequency.

**Example:** python /.../Anonymize/ExtractLeaders.py /.../folderwithpairs



Output

Input



1. **Calibration:**

**Part A)** Runs 10-rep. 5-fold cross-validation for sliding threshold of Positive Predictive Value (PPV) of leading words in output of step 1.

**Input / Parameters:**

-r <int> number of repetitions

-k <int> number of bins in k-fold cross-validations

-l <string> choose any from {„*leaders1*“, „*leaders2*“, „*both*“} to specify, which leading words to use in cross-validation.

-o <string> absolute path of paired documents folder

-ss <string> absolute path of file with street suffixes

-ts <string> absolute path of file with town suffixes

-s <string> absolute path of file with streets

-t <string> absolute path of file with towns

-n <string> absolute path of file with first names

-v verbose output (performance for each threshold)

**Example:** python /.../Anonymize/AnonymizationCrossValidation.py -r 1 -k 10 -l both -o /home/ralf/results/letters/test/test/high\_confidence\_pairs -ss /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/suffixes\_streets.txt -ts /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/suffixes\_villages.txt -s /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/german\_streets.txt -t /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/german\_towns\_and\_villages.txt -n /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/international\_firstnames.txt -d 0.7 0.8 0.9 1.0 -p -f 0.1 0.3 0.5 0.7 0.9

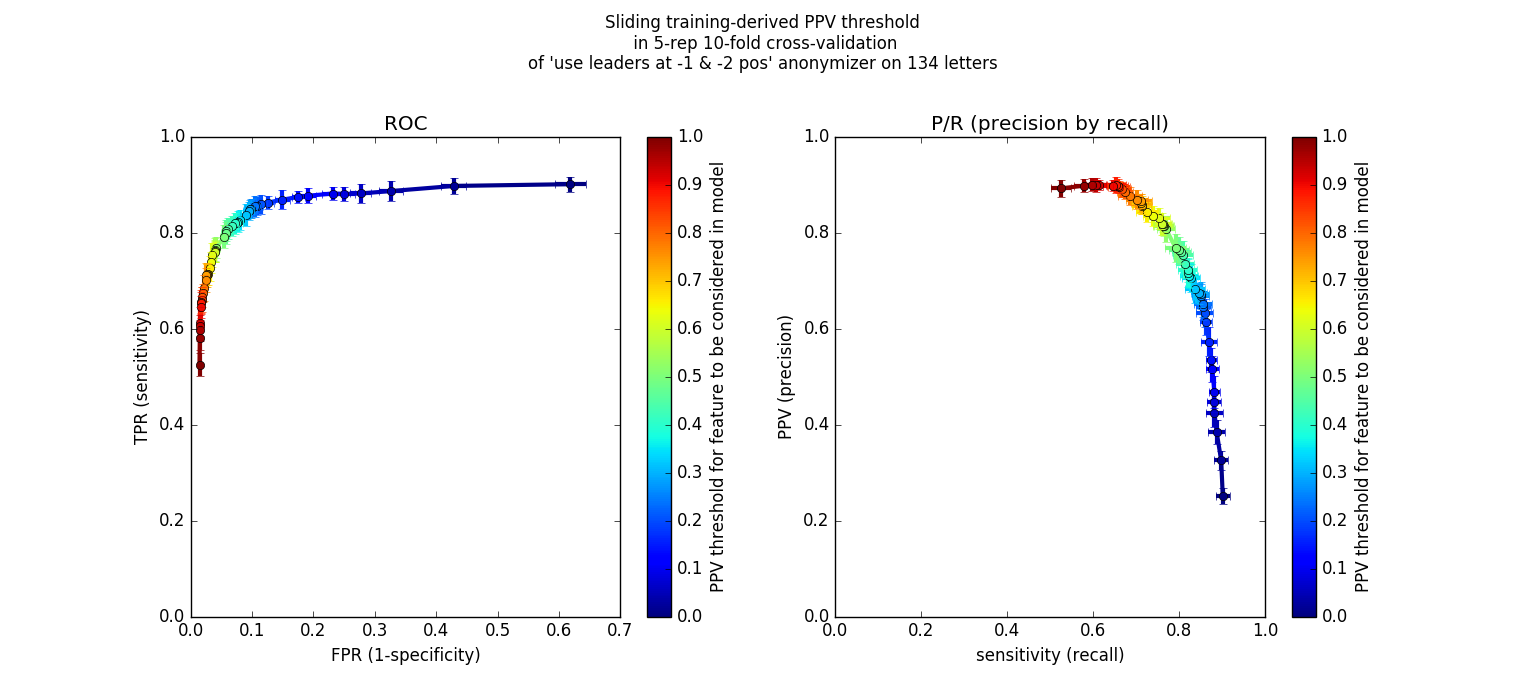
***Output:***

Number of repetitions: 1   
10-fold cross-validation   
Towns to purge: 11651   
Town suffixes to purge: 48   
Streets to purge: 96101   
Street suffixes to purge: 13   
Names to purge: 14356   
Purge emails:  False   
Purge URLs:  False   
1 rep 10 fold CV on 13 pairs of documents ( 10.0 % of data).   
Running cross-validation. Current demasking-threshold: 0.7

**Part B)** Plots Receiver Operating Characteristics (ROC) and Precision / Recall (P/R) curve obtained as model performance under sliding threshold.

**Example:** python /.../Anonymize/PlotLetterROC.py /.../cross\_validation.txt

**Result:**



Choose desired

PPV threshold

1. **Application:** Anonymization of unseen input documents.

**Example:**

python /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/DeidentifyLetters.py -o /home/ralf/results/letters/test/test/random\_sample\_20per\_year -c 0.0 1 0.8 -l /home/ralf/results/letters/test/test/X\_predictors\_1.txt /home/ralf/results/letters/test/test/X\_predictors\_2.txt -ss /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/suffixes\_streets.txt -ts /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/suffixes\_villages.txt -s /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/german\_streets.txt -t /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/german\_towns\_and\_villages.txt -n /home/ralf/Dropbox/Dieterich/Code/anonymization/Anonymize/international\_firstnames.txt -p

**Output:**

Thresholds: {'distinct\_words\_predicted': 1, 'positive\_predictive\_value': 0.8, 'rel\_amount\_of\_documents\_leader\_occurs\_in': 0.0}   
Using  958 -1 leading words.   
Using  937 -2 leading words.   
Expecting target documents to be located in the sub-folders of /home/ralf/results/letters/test/test/random\_sample\_20per\_year   
Sub-folder 2006   
Sub-folder 2016   
Sub-folder 2005   
Sub-folder 2008   
Sub-folder 2012   
Sub-folder 2007   
Sub-folder 2004   
Sub-folder 2010   
Sub-folder 2014   
Sub-folder 2011   
Sub-folder 2009   
Sub-folder 2013

**Result:**

Universitätsklinikum XXXXXXXXXX | XX XXXXXXXXXXX XXXX XXX X XXXXX XXXXXXXXXX Medizinische Klinik (Krehl-Klinik) XXXXXXXXX XXXXXX XXXXXXX XXX Kardiologie XXXXXXXXXX und Pulmologie XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXX XX X X XXXXX Ärztlicher Direktor Station XXXXXXX Im Neuenheimer XXXX XXX XXXXX XXXXXXXXXX Tel XXX XXXXXXXXXXXXXXXXXX Fax XXX XXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXX / XXX XXXX XX XXX XXXX XXXXXXXXX XXXX Kleingemünder Straße XX XXXXX XXXXXXXXXX Nachrichtlich Herrn XXXXXX X XXXXXX 4a Vorläufiger Kurzarztbrief Ein ausführlicher endgültiger Arztbrief folgt Sehr geehrte XXXX XXXXXXXX XXXX wir berichten über Ihre Patientin XXXX XXXXXXXXX XXXXXX geboren am XXXXXXXXXX wh XXXXX XXXXXXXXXX 2 A die sich vom XXX in unserer statio­nären XXXXXXXXXXX befand Diagnosen Aktuell Dilatative Kardiomyopathie mit global mittelgradig eingeschränkter linksventrikulärer Pumpfunktion (HK Nr.123585 vom XXXXXXXXX Koronare Herzerkrankung ohne höhergradige Stenosen Drücke im kleinen Kreislauf im oberen Normbereich Angiographisch kein Hinweis auf eine Mitralklappen- oder Aortenklappeninsuffizienz Synkope,V a neurokardiogene Synkope vom XXXXXXXXXXXXXXXXXXX Typ Paroxysmales Vorhofflimmern Gestörte Glucosentoleranz (HbA1c 5,9%) Diastolische Relaxationsstörung des LV bei konzentrischer Linksherzhypertrophie und enddiastolischer Vergrößerung des LV Mitralklappenprolaps mit begleitender Mitralklappeninsuffizienz II° Aortenklappeninsuffizienz I° Trikuspidalklappeninsuffizienz II° Ventrikuläre Extrasystolie Lown IVa Diffuse..