Unmasking

Identifying Pseudonymous Authors

Winnowing

- N-dimensional boolean feature space
  - X = (0,1,0,0,0,0,1,1,0,1….)
- Model is a n-vector of weights
  - w=(1,1,1,1,1,1,1,1,1,1….)
- Prediction for a new point is dot product
  - $x \cdot w = \text{sum } w_i * x_i > \text{threshold}$
- Training:

```
w = np.ones(n)
#Default weight of one
for x in training_set:
#Pass through points
if (x*w>t)!=y
#False  +  OR  -
w -= (a*x) OR w += (a*x)
#Scale rel. weights by `a`
```

Authorship Attribution

- Categorize new data given examples from two similar authors
- Feature space:
  - 500 most frequent words, minus content (!) words, is 304
    - ‘clear’ ‘question’ ‘opinion’ ‘explicit’ ‘said’ ‘therefore’
- Cross Val:
  - 5-fold division of documents
- 85%-95% accuracy

Authorship Prediction

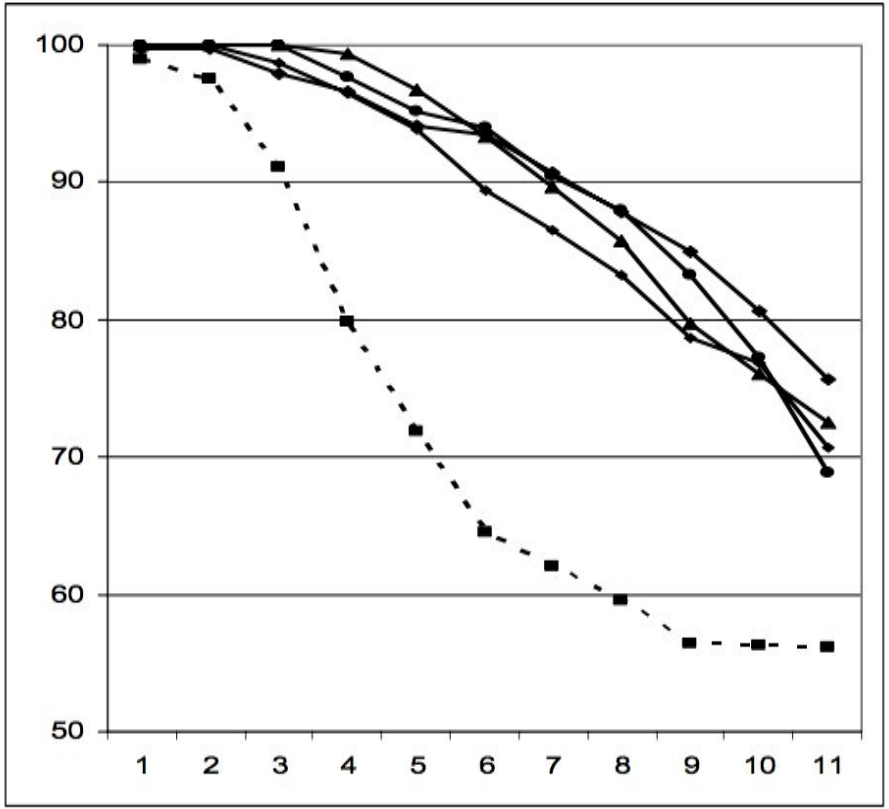
- RP was written by YH.
- We suspect he also wrote TL.
- Take 4 similar works
  - ZZ,SN,DN,GV
- Form pairwise models
  - >95% accuracy

Do the differences between RP and TL indicate different authors or different contexts / styles / genres or even red herrings?

Authorship Unmasking

- Form pairwise models to TL
- Remove 5 highest features
- Repeat

See how much quicker TL and RP become indistinguishable



Authorship Credit

M. Koppel, D. Mughaz and N. Akiva (2006), "New Methods for Attribution...", *Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics*

Thank you!