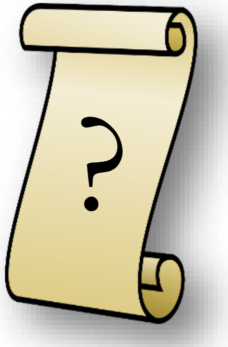
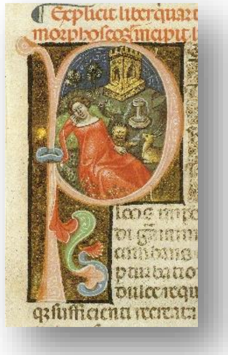
Determining If Two Documents Are By The Same Author

Moshe Koppel and Yaron Winter

Open-Set Author Verification Problem

Identify if (potentially short) texts X and Y are by the same author

( , ) = ✓ / ✗

Explicit liber quart
morphoseos incipit l

Running Example: *Compendiosa expositio*

Stover, Winter, Koppel, & Kestemont 2016

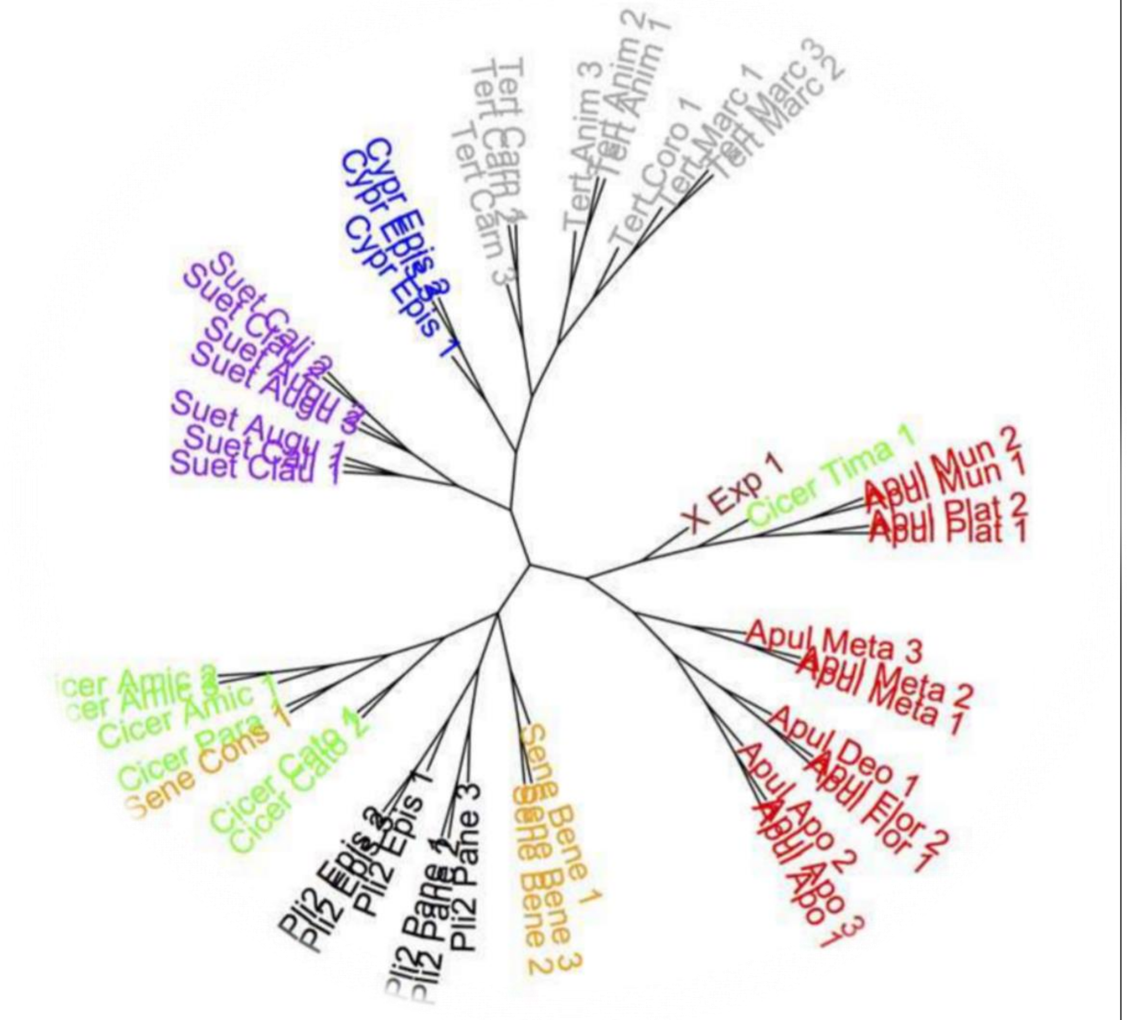
- Single medieval manuscript in the Vatican Library in Rome
- Philological analysis indicated that the text is likely from antiquity
- Traditional stylistic and metrical analyses suggest the author is Apuleius of Maudoros
- Goal: Verify that the *Expositio* was written by Apuleius

Compendiosa expositio

- It clusters with the works by Apuleius

- **Problem**: Clustering isn’t perfect

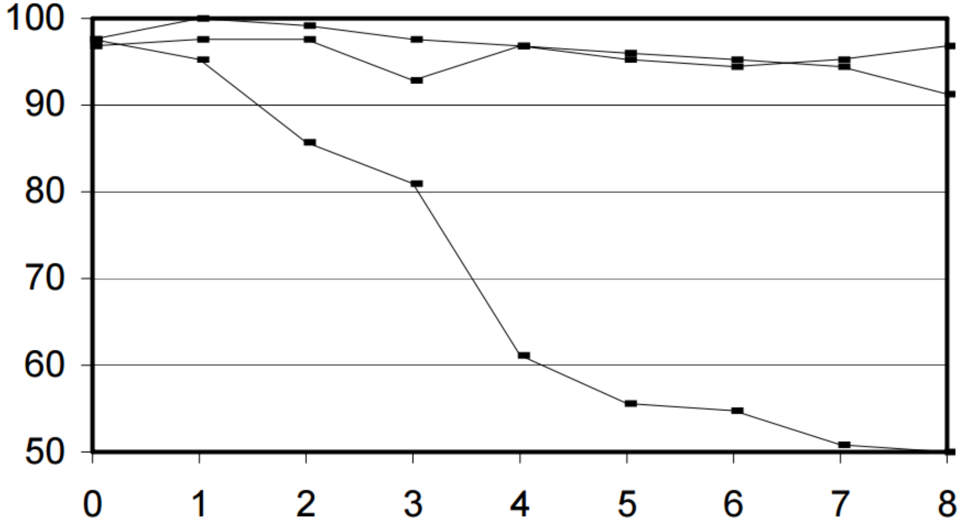
- Need verification because we cannot assume that the true author is among the available candidates



Related Work: Unmasking Method

Koppel & Schler 2004

- Idea: If books X and Y are by the same author, then their differences are reflected in only a small number of features
- "Unmasking" = Iteratively remove most distinguishing features and see how quickly cross-validation accuracy degrades



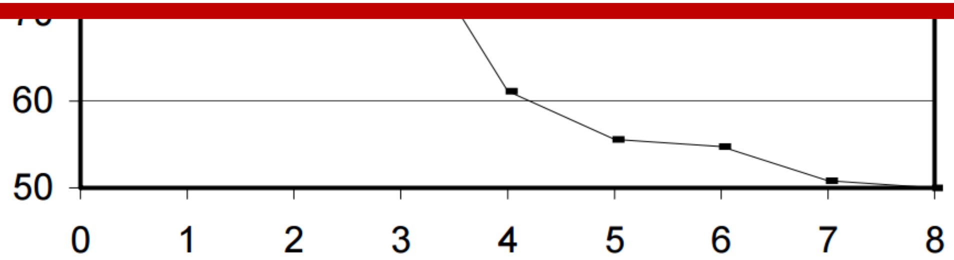
Ten fold cross-validation accuracy of models distinguishing *House of Seven Gables* from each of Hawthorne, Melville and Cooper. The x-axis represents the number of iterations of eliminating best features at previous iteration.

Related Work: Unmasking Method

Koppel & Schler 2004

Problem: Method relies on paper chunking.

Ineffective for short input documents (< 10,000 words)
Sanderson & Guenter 2006



axis represents the number of iterations of eliminating best features at previous iteration.

Related Work: Authorship Attribution

Type I: Machine-Learning Methods

- Idea: For each candidate author, construct a classifier from their literary works

- Abbasi & Chen 2008
- Koppel, Schler, & Argamon 2008
- Zhao & Zobel 2005
- Zheng Li, Chen, & Huang 2006

- **Problem:** Does not scale well with large number of possible authors

Related Work: Authorship Attribution

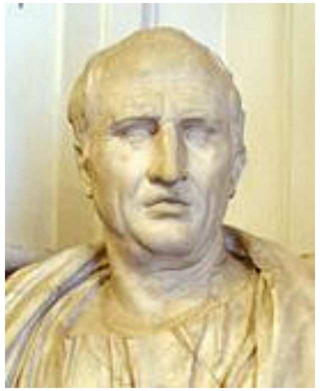
Type II: Similarity-Based Methods

- Idea: Measure the "distance" between two documents. Attribution is given to the author with the closest corpus (one collective document)

- Abbasi & Chen 2008
- Argamon 2007
- Brennan & Greenstadt 2009
- Burrows 2002
- Hoover 2003
- Malyutov 2006
- Uzuner & Katz 2006

The Many-Candidates Problem

- Open-set identification problem
- Given a large set of candidates determine which, if any, of them is the author of a given anonymous document



The Many Candidates Method

Koppel, Schler, & Argamon 2011

```
Given: A snippet to be assigned; known-texts for each of C candidates
1.  Repeat k times
    a.  Randomly choose half of the features in the full feature set
    b.  Find top known-text match to snippet using min-max similarity
2.  For each candidate author A,
    a.  Score(A) = proportion of times A is top match
Output: $\text{argmax}_A$ Score(A) if max Score(A) > $\sigma^*$; else Don't Know
```

The Impostors Method

- Can convert the verification problem into the many-candidates problem by generating a large set of impostor candidates
- Well-established practice in the speaker-identification community
- Method of impostor generation is important
  - Too few or unconvincing impostors will produce too many false positives
  - Too many impostors or genre imbalance will produce too many false negatives

The Impostors Method

1. Generate a set of impostors $Y_1, \dots, Y_m$
2. Compute $score_X(Y)$ = the number of choices of feature sets (out of 100) for which $sim(X, Y) > sim(X, Y_i)$, for all $i = 1, \dots, m$
3. Repeat the above with impostors $X_1, \dots, X_m$ and compute $score_Y(X)$ in an analogous manner
4. If $\text{avg}\big(score_X(Y), score_Y(X)\big) > \sigma^*$, assign $\langle X, Y \rangle$ to same-author

Experimental Setup

- Universe: All blogs by several thousand bloggers from blogger.com
  - On average, 38 separate blog posts per author
- Consider pairs of fragments of blog posts: $\langle X, Y \rangle$
  - X = First 500 words produced by a given blogger
  - Y = Last 500 words produced by a given blogger
- Corpus: 500 pairs such that 250 are same-author and 250 are not
  - No single blogger appears in more than one pair

Similarity-Based Baseline Method

- Measure the similarity between X and Y and label the pair as same-author when the similarity exceeds some threshold σ^*
- Represent X and Y as vectors such that each entry represents the tf-idf value of a character 4-gram of the corresponding document
- Similarity Measures:

1. Cosine: $\text{sim}(X,Y) = \cos\left(\vec{X},\vec{Y}\right) = \frac{\vec{X}*\vec{Y}}{\|\vec{X}\|*\|\vec{Y}\|}$
2. Min-Max: $\text{sim}(X,Y) = \text{minmax}\left(\vec{X},\vec{Y}\right) = \frac{\sum_{i=1}^{n}\min(x_i,y_i)}{\sum_{i=1}^{n}\max(x_i,y_i)}$

Similarity-Based Baseline Method

Development Set Accuracy:

- Cosine: 70.6%
- Minmax: 74.2%

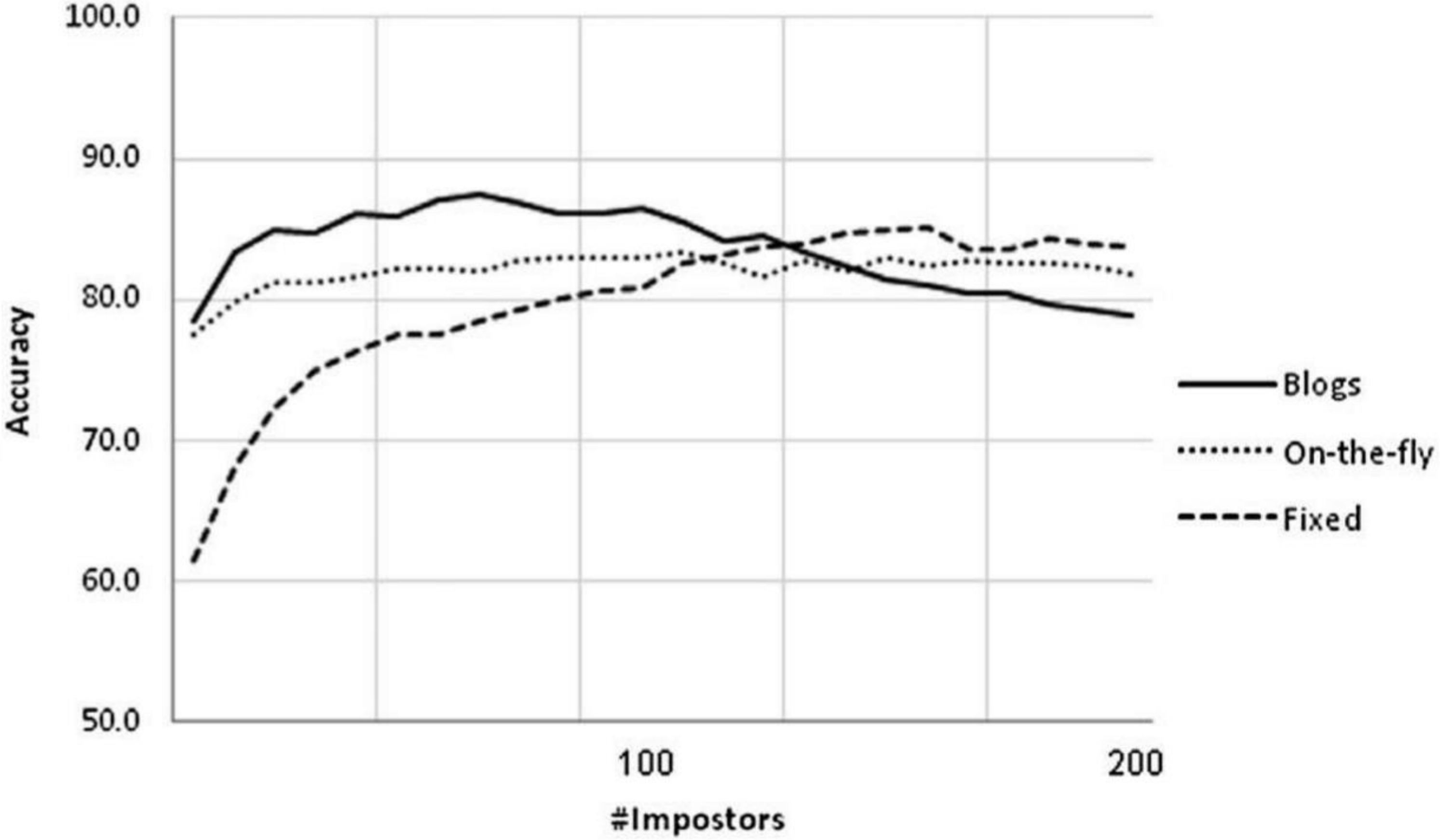
1. Cosine: $\text{sim}(X, Y) = \cos\left(\vec{X}, \vec{Y}\right) = \frac{\vec{X} * \vec{Y}}{\|\vec{X}\| * \|\vec{Y}\|}$
2. Min-Max: $\text{sim}(X, Y) = \text{minmax}\left(\vec{X}, \vec{Y}\right) = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)}$

Supervised Baseline Method

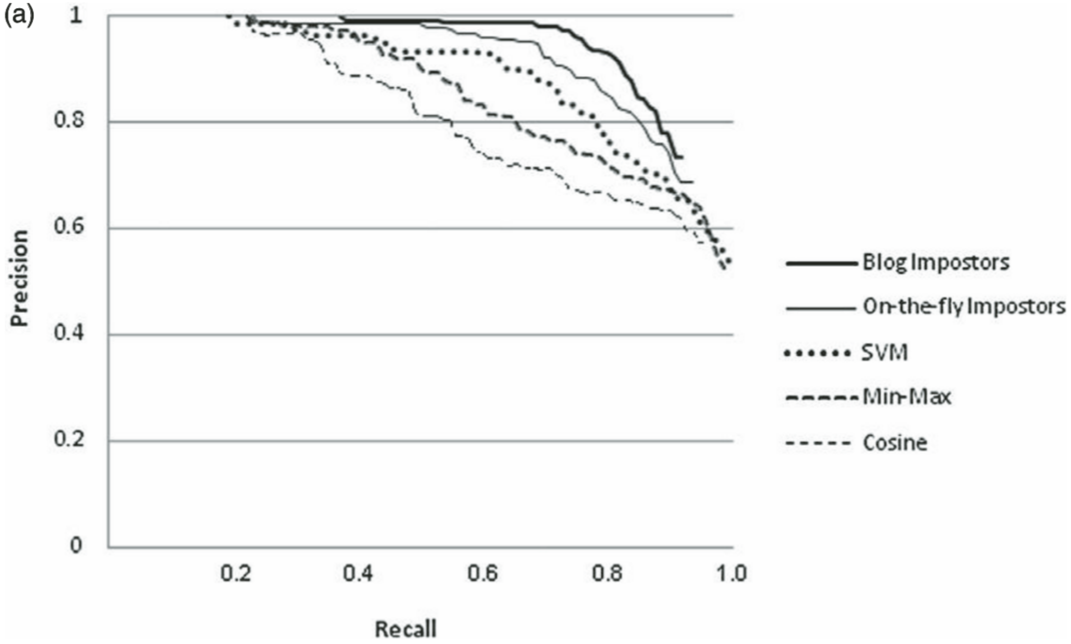
- Training set: 1,000 labelled $\langle X, Y \rangle$
  - Train on labelled difference vectors: $\text{diff}(X, Y) = \left|\vec{X} - \vec{Y}\right|$
- Learn a linear SVM classifier from the labelled vectors
  - Learns nothing about specific authors, only what differences in n-gram frequencies characterize same-author pairs in general

Development Set Accuracy: 79.8%

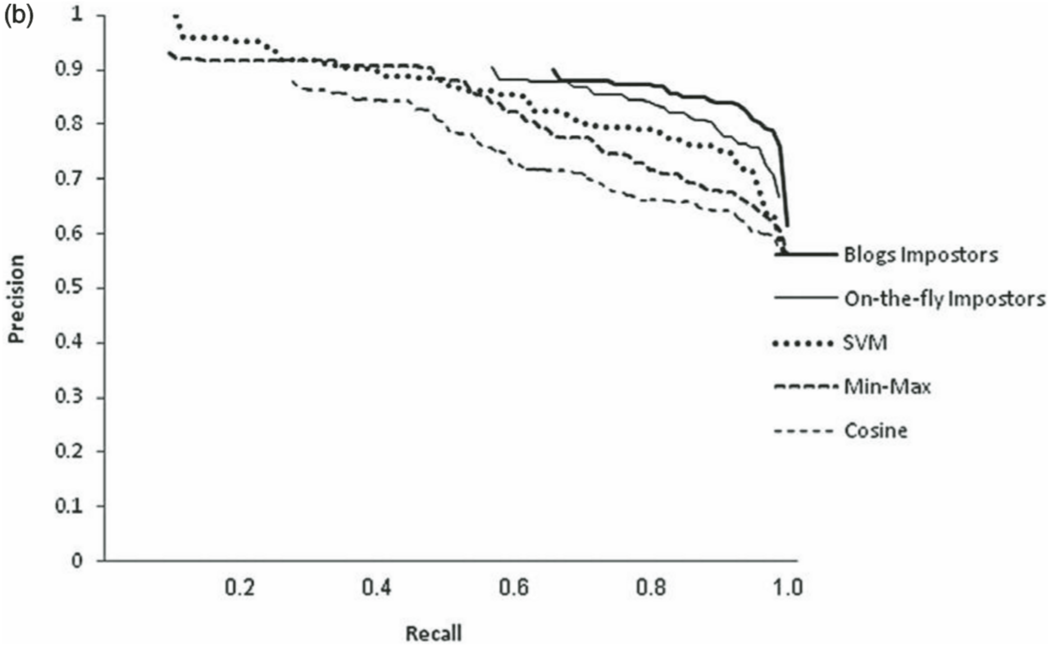
Impostor Generation



Results

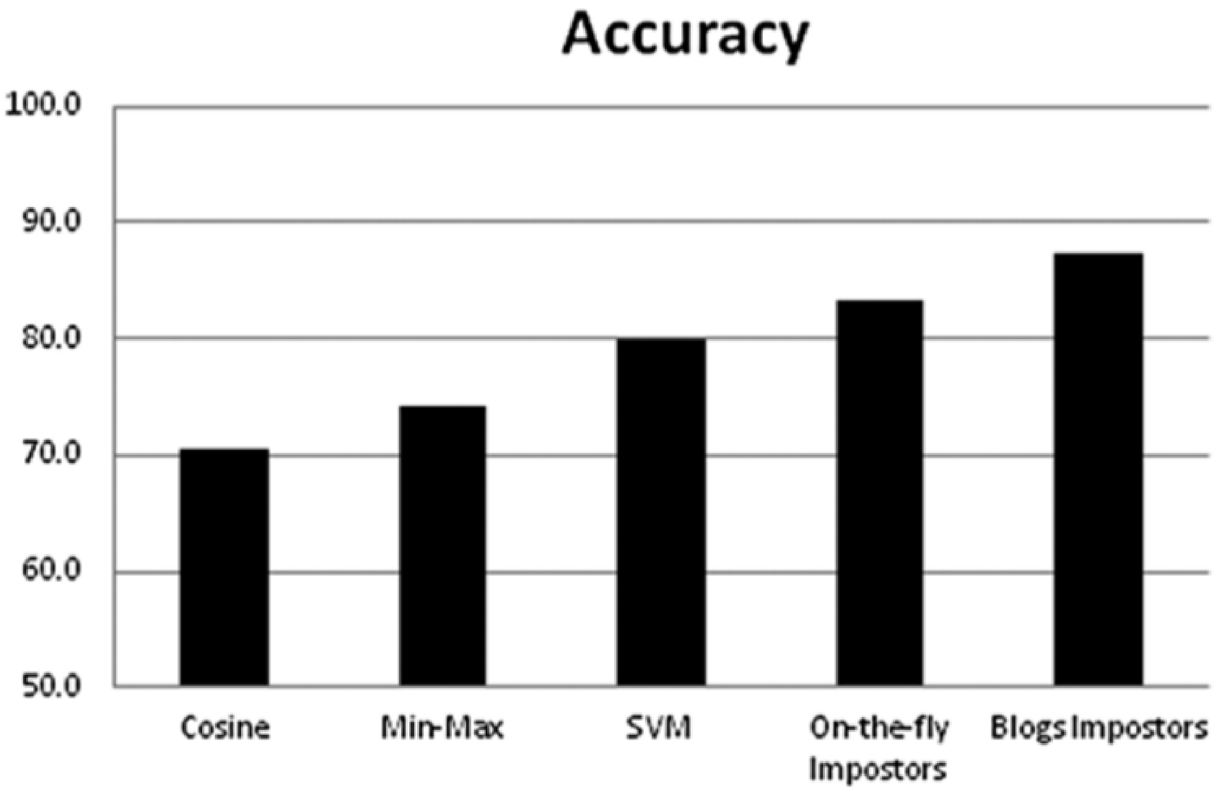


Same-Author



Different-Author

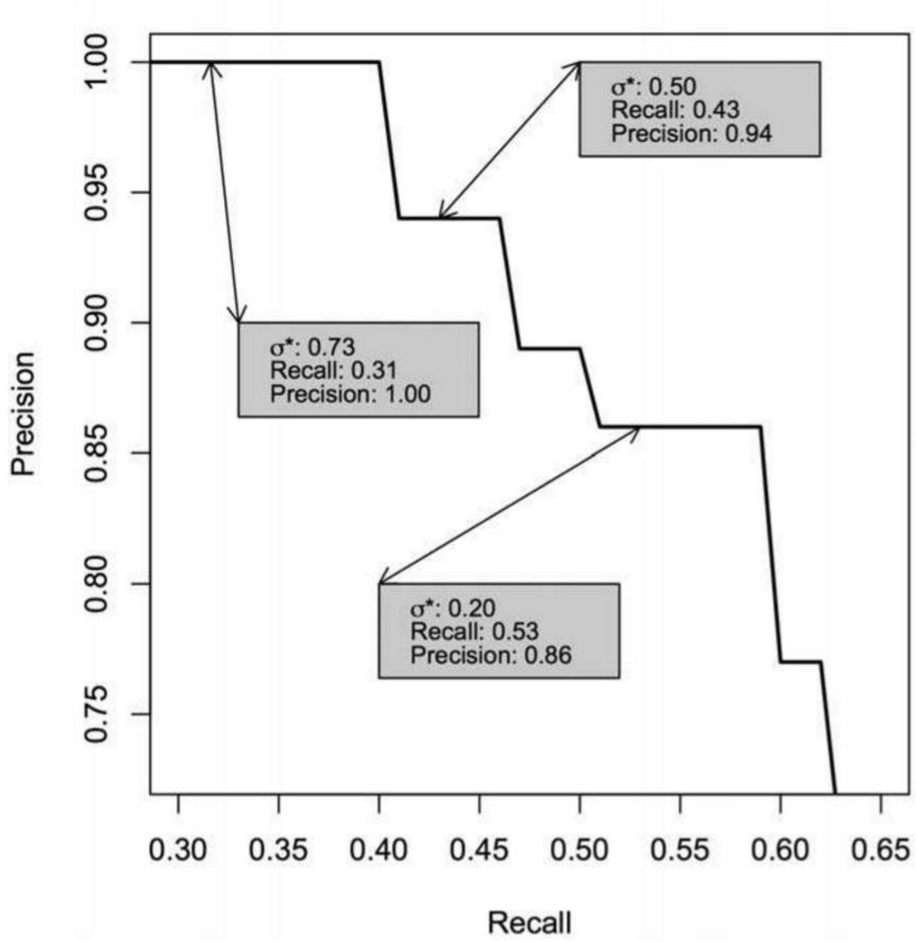
Results


Accuracy
100.0
90.0
80.0
70.0
60.0
50.0
Cosine
Min-Max
SVM
On-the-fly Impostors
Blogs Impostors

Experimental Setup: *Compendiosa expositio*

- Development Corpus: 22 texts by authors with stylistic, chronological, generic, or thematic similarity with Apuleius
- Random Feature Set: 125,000 of 250,000 unigrams and bigrams
- Background Set: 180 texts by 36 authors writing in similar genres and/or periods
- Imposter Set: 50 texts randomly selected from background set

Results: *Compendiosa expositio*



Precision-Recall curve. The effect of various thresholds σ^* for the verification score in terms of precision and recall for the same-author category in the development corpus

Results: *Compendiosa expositio*

- Similarity measures for Apuleius's work:
  - *De deo Socratis*, *Florida*, and *Apologia* have a score of 0.85+
  - *Metamorphoses* (*The Golden Ass*) has a score above 0.50 with **only** *Florida*
  - No pairings of Apuleius's works with other texts surpassed 0.35
- "Non-greedy" attributor
  - High precision, but relatively low recall for same-author pairs
- A new text X would be extremely likely to have been written by Apuleius if $\langle X, Y_{Apuleius} \rangle$ obtains a score above 0.20

Results: *Compendiosa expositio*

- The pair *Expositio* and *De Platone* has a score of 0.73
- No other text pairing with the *Expositio* has a score above 0.04
- Lends support to the hypothesis that the *Expositio* is the forgotten third book of *De Platone*
- These results emphasize the importance of genre
  - The *Expositio*'s genre of Platonic philosophy matches the *De Platone*, but does not match the majority of Apuleius's work

Limitations

- If documents X and Y are in different genres, it is much more difficult to distinguish same-author/different-author pairs
- Need strong confidence that impostor documents are not written by the authors of documents X and Y

Conclusion

- Introduce an almost unsupervised approach for determining if a pair of short documents is written by the same author
    - Two phases:
        1. Generate impostor set
        2. Use feature randomization to iteratively measure document pair similarity
- There is a fine balance between impostor quality and quantity
    - The better the impostors, the fewer are needed
- Corroborate that Apuleius wrote the *Compendiosa expositio*