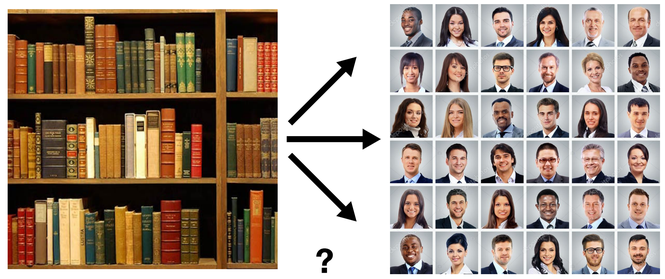
Authorship Identification with Support Vector Machines

Austin Hoover
University of Tennessee, Knoxville

Overview

Basic task: match each document with the correct author



Dataset

- 5000 documents, 50 authors
- Average document length < 4 paragraphs
- All authors share common subject area

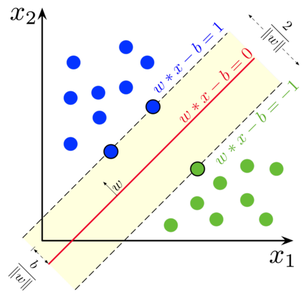
Character n-grams

a_string → (n=4) → a_st | _str | stri | trin | ring

Tools

Support Vector Machine (SVM)

- Find optimal separating plane
- Scales to arbitrary number of dimensions
- For high dimensionality, points are usually linearly separable



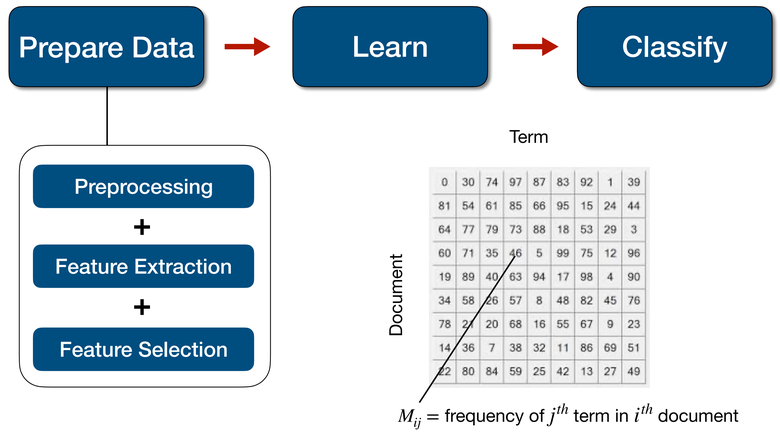
Multi-class classification

- *One-vs-all:* train n_{class} classifiers
- For each document, choose class with "best" separating plane.

Feature Selection Methods

- Mutual Information — $I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$
- χ^2
- Anova F-value

Method



Questions to answer:

- Which n-gram length is best: 3, 4, 5, or multi-length?
- Which feature selection method is best?
- What is the optimal number of features?
- Can the performance be improved?
  - Scaling, preprocessing, etc.

Results

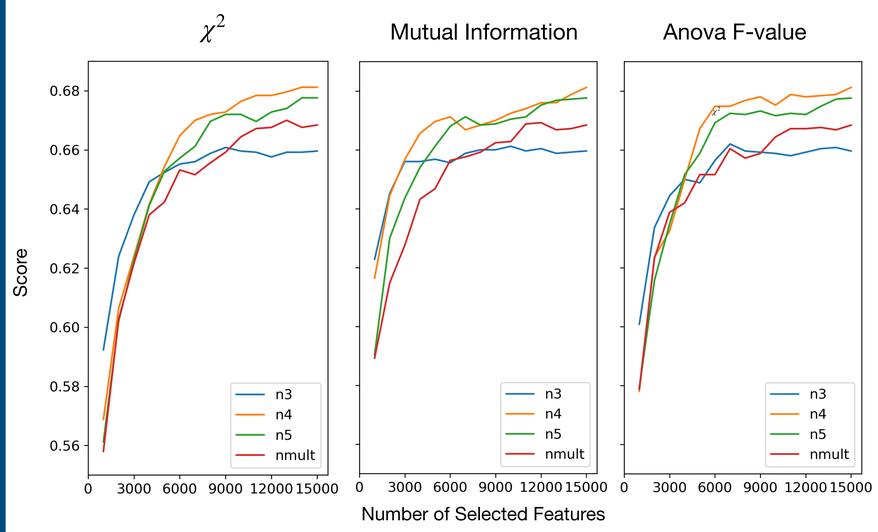


Fig 1: *Accuracy vs number of selected features for 3 selection methods.*

The original feature set is the 15,000 most frequent 3/4/5 grams, as well as equal parts 3/4/5 grams. 3-grams perform well at low feature numbers, while 4-grams are best as the dimensionality increases. In all cases, information is lost and accuracy decreases when the entire feature set is not considered. Above 15,000 features, the performance increase is negligible. Mutual information (center figure) is superior to the neighboring methods.

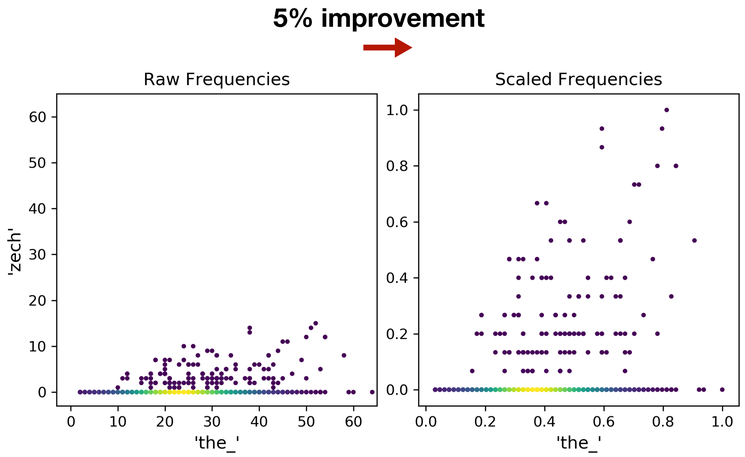


Fig 2: *N-gram frequencies for "the_" and "zech".*

Scaling all ngram counts to the range [0,1] results in a 5% increase in accuracy score. In the dataset, one author writes frequently about the Czech Republic. In the figure it is apparent that after scaling, occurrences of the uncommon string *zech* will be more heavily weighted.

Conclusions

- **Optimal feature set**
  - Mutual information is best performing feature selection method
  - 4-grams outperform 3, 5, and multi-length n-grams
  - Accuracy ~ log (number of features).
- **Effect of scaling frequencies**
  - Scaling features improves accuracy
  - Max accuracy ~ 75%
- **Possible extensions**
  - The multi-length n-grams extracted contain duplicate strings. For example: *the, the_, _the, _the_.* Alternate feature selection methods take neighboring n-grams into account and can achieve better performance at low feature numbers [1].

References

[1] Houvardas, John and Efstathios Stamatatos. *N-Gram Feature Selection for Authorship Identification.* AIMSA (2006).

[2] Holmes, D.: *The Evolution of Stylometry in Humanities Scholarship.* Literary and Linguistic Computing, 13:3 (1998) 111-117.

[3] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. *Authorship Attribution with Support Vector Machines.* Applied Intelligence 19, 1-2 (May 2003), 109-123. DOI:https://doi.org/10.1023/A:1023824908771