EXPLORING THE AUTHOR VERIFICATION (AV) TASK

T. Pace & S. Nystrom [1]

[1] *School of Computing, University of Utah*

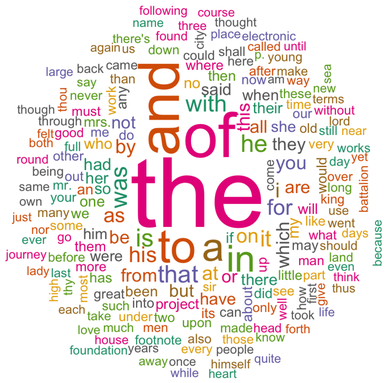
WHY AV?

- A “stylistic fingerprint”
- Dynamic area of research!
- Extensive practical applications
  - (books, various online platforms)

OUR PROCESS

1. Develop a web scraper to download e-books
   - Project Gutenberg.com
2. Calculate 12 distinct features & TF-IDF
3. Dimensionality reduction (PCA & LDA)
4. Cluster analysis
5. Evaluate performance (via Shannon entropy)

DO STOP WORDS MATTER?



Now let’s see the results!

We find that **cluster analysis** that uses **TF-IDF** outperforms other feature selections.

THE UNIVERSITY OF UTAH

RESULTS

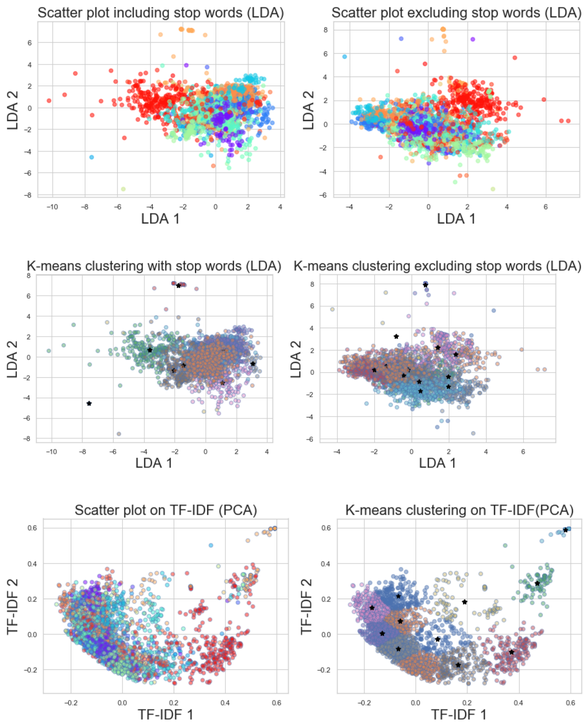


Figure 1: Scatter plots vs. K-means clustering for our standard features and TF-IDF

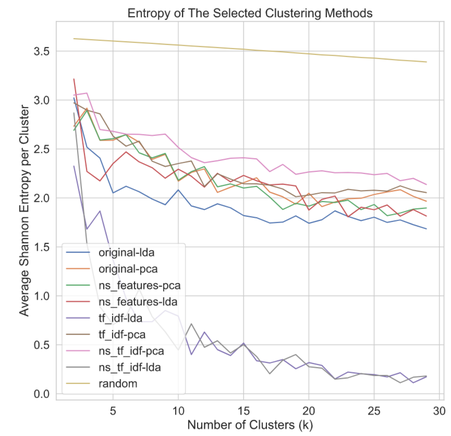


Figure 2: Shannon entropy for the selected clustering methods and values of k, the number of clusters

TAKE-AWAYS & THANK YOUS

- TF-IDF using LDA performed the best!
- Little difference when *stop words* were included/excluded for our original features
- Need for **MORE** separation *between* clusters
- Thank you for viewing our poster!