Paragraph Clustering for Intrinsic Plagiarism Detection

Using a Stylistic Vector Space Model with Extrinsic Features

Julian Brooke and Graeme Hirst
University of Toronto

Model

- Based on work on detecting stylistic inconsistency
  - In particular, voice segmentation in poetry
- Extrinsic features from lexicons and larger corpora
  - Using LSA to derive a stylistic lexicon
- Cluster by maximizing distance between authors
  - Correct for imbalance in span size using expected difference between sums of random variables

Results

- Development evaluation
  - Method works well overall
  - Correcting for span difference is important
- But poor performance in PAN multi-author evaluation on mixed novels
  - Model is too conservative
  - Few stylistic differences between novels
  - Major stylistic differences between dialogue and narration, which confuses model