Authorship Classification

Problem Definition

- Authorship classification, the science of inferring characteristics of the author from the characteristics of documents written by that author, is a problem with a long history and a wide range of application.
-
-

Data Parse

- Word tokenization
- Sentence tokenization
- Stemming
- Label part of speech tag
-

Features Sets

Stylometry
Vocabulary Diversity
Bags of Words
Frequency of function words (particle, pronoun, conjunction)

Features

- Stylistic Features

Number of Sentences	Number of words	Average Sentence Length
Average Word Length	Number of Different Words	Number of Periods
Number of Commas	Number of Colons	Number of Semi colons
Number of Exclamation Marks	Number of Question Marks	

Features

- Bag of Words
  - The bag-of-words representation, where the document is represented with a vector of the word counts that appear in it. Depending on the classification method, the bag-of-words vector can be normalized to unity and scaled so that common words are less important than rare words

Methods and Testing

- Bayes Classifier
  - For stylometry feature set.
    - Covariance matrix for each class (each writer is a different class)
  - For function words feature set.
    - Bayes Classifier with Gaussian Density
- SVM Classifier
  - For bag of words
    -

Baseline

- 90% accuracy reported by Peng et al. (2003) - n-gram model
- 95% accuracy reported by Ge et al. (2016) - NNLM(neural network language model)

Evaluation Metric

- The performance of an authorship classifier can be naturally measured by its overall accuracy: the number of correctly classified texts divided by the number of texts classified overall.

- For each categories, we use precision, recall and F-measure

Future Work

- Complete proposed experiment
- Compare with the baseline
- Consider improvement of feature selection and algorithm

Reference

- Peng, F.; Schuurmans, D.; Keselj, V; Wang, S. (2003). Language Independent Authorship Attribution using Character Level Language Models. Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, 267--274
- A. (Ed.). (2015). Authorship Attribution with Python. Retrieved November 15, 2016, from Piazza.com/class/isg9se99b5s6nw?cid=1
- Nirkhi, S. (2015, May). Stylometric Approach For Author Identification of Online Messages. 1-2. Retrieved from http://ijcsit.com/docs/Volume 5/vol5issue05/ijcsit2014050536.pdf
- K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott,"Stylometry for E-mail Author Identification and Authentication",Seidenberg School of CSIS, Pace University, New York
- ZhiLiu, Reuter_50_50 Data Set. Retrieved November 15, 2016, from http://archive.ics.uci.edu/ml/datasets/Reuter_50_50
- Shihara, S. (n.d.). A Forensic Authorship Classification in SMS Messages: A Likelihood Ratio Based Approach Using N-gram. *Likelihood Ratio Calculation,* 51-53. Retrieved from http://www.alta.asn.au/events/alta2011/proceedings/pdf/U11-1008.pdf
- Castro, A., & Lindauer, B. (n.d.). Author Identification on Twitter. Retrieved November 15, 2016, from http://cs229.stanford.edu/proj2012/CastroLindauer-AuthorIdentificationOnTwitter.pdf
- Elayidom, S., Jose, C., Puthussery, A., & Sasi, N. K. (2013, September 5). TEXT CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION ANALYSIS. 1-8. Retrieved November 14, 2016, from http://airccse.org/journal/acij/papers/4513acij01.pdf
- Nirkhi, S., Dharaskar, R., & Thakare, V. (n.d.). Authorship Identification in Digital Forensics using Machine Learning Approach. Retrieved from http://www.ijltet.org/wp-content/uploads/2015/01/54.pdf
- AICBT, Authorship Attribution with Python. Retrieved November 8, 2016, from http://www.aicbt.com/authorship-attribution/
- Nazar, R; Marta, S.P. (2006) An Extremely Simple Authorship Attribution System. Proceedings of the Second European IAFL Conference on Forensic Linguistics / Language and the Law, Barcelona 2006.