2017
Certificate Authority Cup International Mathematical Contest Modeling
http://mcm.tzmcm.cn

Problem B(MCM): Handwriting Analysis in the Email

Handwriting analysis is a highly specific form of investigation that is used to link people to written evidence. Handwriting investigators are generally called upon in a court of law or criminal investigation to identify whether or not a writing sample comes from a particular person. Since many language evidence now appears in e-mail, in a broad sense, handwriting analysis also includes the problem of how to identify the author by the linguistic features of the Email.

Authorship attribution is the process in which linguists set out to identify the author(s) of disputed texts using identifiable features of linguistic style, ranging from word frequencies to preferred syntactic structures. The content of Email tends to be short and the author's linguistic style is quite obvious. Please structure an effective model to identify the author through capturing the linguistic features of the Email. You can use the Enron Email Datasets to train and test your model.

The Enron Email Datasets link: http://bailando.sims.berkeley.edu/enron_Email.html