

Table of contents

Preface

IX

CHAPTER 1

Author identification

1

1. Introduction 1
2. Feature selection 5
 - 2.1 Evaluation of feature sets for authorship attribution 8
3. Inter-textual distances 11
 - 3.1 Manhattan distance and Euclidean distance 12
 - 3.2 Labbé and Labbé's measure 14
 - 3.3 Chi-squared distance 15
 - 3.4 The cosine similarity measure 16
 - 3.5 Kullback-Leibler Divergence (KLD) 18
 - 3.6 Burrows' Delta 18
 - 3.7 Evaluation of feature-based measures for inter-textual distance 23
 - 3.8 Inter-textual distance by semantic similarity 26
 - 3.9 Stemmatology as a measure of inter-textual distance 28
4. Clustering techniques 30
 - 4.1 Introduction to factor analysis 31
 - 4.2 Matrix algebra 35
 - 4.3 Use of matrix algebra for PCA 38
 - 4.4 PCA case studies 44
 - 4.5 Correspondence analysis 45
5. Comparisons of classifiers 47
6. Other tasks related to authorship 50
 - 6.1 Stylochronometry 50
 - 6.2 Affect dictionaries and psychological profiling 53
 - 6.3 Evaluation of author profiling 58
7. Conclusion 58

CHAPTER 2

Plagiarism and spam filtering

59

1. Introduction 59
2. Plagiarism detection software 62
 - 2.1 Collusion and plagiarism, external and intrinsic 63
 - 2.2 Preprocessing of corpora and feature extraction 63
 - 2.3 Sequence comparison and exact match 64
 - 2.4 Source-suspicious document similarity measures 65
 - 2.5 Fingerprinting 66
 - 2.6 Language models 67
 - 2.7 Natural language processing 68
 - 2.8 Intrinsic plagiarism detection 70
 - 2.9 Plagiarism of program code 73
 - 2.10 Distance between translated and original text 74
 - 2.11 Direction of plagiarism 76
 - 2.12 The search engine-based approach used at PAN-13 78
 - 2.13 Case study 1: Hidden influences from printed sources in the Gaelic tales of Duncan and Neil MacDonald 81
 - 2.14 Case study 2: General George Pickett and related writings 83
 - 2.15 Evaluation methods 84
 - 2.16 Conclusion 85
3. Spam filters 86
 - 3.1 Content-based techniques 87
 - 3.2 Building a labeled corpus for training 87
 - 3.3 Exact matching techniques 88
 - 3.4 Rule-based methods 89
 - 3.5 Machine learning 90
 - 3.6 Unsupervised machine learning approaches 92
 - 3.7 Other spam-filtering problems 93
 - 3.8 Evaluation of spam filters 94
 - 3.9 Non-linguistic techniques 94
 - 3.10 Conclusion 97
4. Recommendations for further reading 98

CHAPTER 3

Computer studies of Shakespearean authorship

99

1. Introduction 99
2. Shakespeare, Wilkins and “Pericles” 101
 - 2.1 Correspondence analysis for “Pericles” and related texts 105
3. Shakespeare, Fletcher and “The Two Noble Kinsmen” 108
4. “King John” 110

5. "The Raigne of King Edward III" 111
 - 5.1 Neural networks in stylometry 111
 - 5.2 Cusum charts in stylometry 113
 - 5.3 Burrows' Zeta and Iota 116
6. Hand D in "Sir Thomas More" 118
 - 6.1 Elliott, Valenza and the Earl of Oxford 118
 - 6.2 Elliott and Valenza: Hand D 121
 - 6.3 Bayesian approach to questions of Shakespearian authorship 122
 - 6.4 Bayesian analysis of Shakespeare's second person pronouns 127
 - 6.5 Vocabulary differences, LDA and the authorship of Hand D 130
 - 6.6 Hand D: Conclusions 131
7. The three parts of "Henry VI" 132
8. "Timon of Athens" 132
9. "The Puritan" and "A Yorkshire Tragedy" 133
10. "Arden of Faversham" 134
11. Estimation of the extent of Shakespeare's vocabulary and the authorship of the "Taylor" poem 136
12. The chronology of Shakespeare 141
13. Conclusion 147

CHAPTER 4

Stylometric analysis of religious texts 149

1. Introduction 149
 - 1.1 Overview of the New Testament by correspondence analysis 151
 - 1.2 Q 153
 - 1.3 Luke and Acts 169
 - 1.4 Recent approaches to New Testament stylometry 171
 - 1.5 The Pauline Epistles 175
 - 1.6 Hebrews 188
 - 1.7 The Signs Gospel 188
2. Stylometric analysis of the Book of Mormon 190
3. Stylometric studies of the Qu'ran 198
4. Conclusion 206

CHAPTER 5

Computers and decipherment 207

1. Introduction 207
 - 1.1 Differences between cryptography and decipherment 208
 - 1.2 Cryptological techniques for automatic language recognition 209
 - 1.3 Dictionary approaches to language recognition 212
 - 1.4 Sinkov's test 212

1.5	Index of coincidence	213
1.6	The log-likelihood ratio	214
1.7	The chi-squared test statistic	215
1.8	Entropy of language	215
1.9	Zipf's Law and Heaps' Law coefficients	218
1.10	Modal token length	219
1.11	Autocorrelation analysis	220
1.12	Vowel identification	221
2.	Rongorongo	224
2.1	History of Rongorongo	224
2.2	Characteristics of Rongorongo	226
2.3	Obstacles to decipherment	227
2.4	Encoding of Rongorongo symbols	227
2.5	The "Mamari" lunar calendar	228
2.6	Basic statistics of the Rongorongo corpus	228
2.7	Alignment of the Rongorongo corpus	229
2.8	A concordance for Rongorongo	231
2.9	Collocations and collocations	233
2.10	Classification by genre	234
2.11	Vocabulary richness	237
2.12	Podzniakov's approach to matching frequency curves	241
3.	The Indus Valley texts	243
3.1	Why decipherment of the Indus texts is difficult	243
3.2	Are the Indus texts writing?	244
3.3	Other evidence for the Indus Script being writing	248
3.4	Determining the order of the Markov model	248
3.5	Missing symbols	249
3.6	Text segmentation and the log-likelihood measure	249
3.7	Network analysis of the Indus Signs	251
4.	Linear A	252
5.	The Phaistos disk	255
6.	Iron Age Pictish symbols	256
7.	Mayan glyphs	256
8.	Conclusion	257

References

259

Index

281