Contents

1 Introductory Material 1

1.1 The Demand for Social Media Learning 2
1.2 Business Perspective of Geolocalization 3
1.3 Unveiling the Power of Topics Modeling 4
1.4 Summary of Problems Addressed 5
1.5 Main Research Contributions 5
1.6 Publications 6
1.7 Chapter Overview 7

2 Technical Basics 9

2.1 Overview of Machine Learning 10
2.2 Supervised Learning 10
2.3 Supervised Learning Evaluation 11
2.4 Unsupervised Learning 11
2.5 K-means Clustering 12
2.6 K-nearest Neighbor 13
2.7 Support Vector Machines 14
2.8 Matrix Factorization 16
2.9 Nonnegative Matrix Factorization 16
2.10 Data Pre-Processing 18
2.10.1 Unigram tokenization 18
2.10.2 Bag-of-words representation 18
2.11 Introduction to TFIDF 18
2.12 Filtering Stopwords in a Tokenized Document 19

3 Problem Description 21

3.1 Geolocalization 22
3.2 Conceptual Stability Based Topic Modeling 23
3.3 General Framework for Discovering Hierarchy of Topics from Twitter Data 24

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

4 Datasets **27**
4.1 Textual Data 28
4.2 Geolocation Dataset for User Geolocation Scenario 28
4.3 Geolocation Datasets for real-time Geolocation Scenario 29
4.4 Topic Modeling Datasets 29
4.5 Twitter Dataset for Hierarchical Topic Discovery 31

5 User Geolocalization **33**
5.1 Introduction 34
5.2 Brief Review of Compared Baselines 34
5.3 Geolocation via Clustering Based Discretization 35
5.3.1 Problem Formulation and Notation 35
5.3.2 Clustering Based Discretization Method 36
5.3.2.1 K-means clustering 36
5.3.2.2 Support Vector Machines 36
5.3.2.3 K-nearest Neighbor 37
5.4 Dataset 38
5.5 Evaluation Metrics 38
5.6 Results and Comparison 38
5.7 Hyperparameters Setup 40
5.8 Reproducibility 40

6 Real-time Geolocalization **41**
6.1 Why is Real-time Scenario Important? 42
6.2 Geolocation Revisited and the new Scenario 42
6.3 Lack of Real-time Geolocation Research in Literature 43
6.4 Background and Notation 45
6.5 The Geo Matrix Factorization Model 45
6.6 Model fitting 46
6.7 Inference for test data 49
6.8 Hyperparameters setup 50
6.9 Reproducibility 52
6.10 Results and Discussion 53

7 Conceptual Stability Analysis **55**
7.1 Introduction 56
7.1.1 Clustering Solutions via Nonnegative Matrix Factorization 56
7.1.2 Selection of the Desired Low-rank K 57
7.2 Background on State-of-the-art Research Directions 57
7.2.1 Description on Consensus Clustering 58
7.2.2 Description on Stability Analysis 58
7.3 State-of-the-art NMF-based Methods 58
7.3.1 The Selection of the Number of Topics via Consensus Matrix . . 59

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

7.3.2 The Selection of the Number of Topics via Stability Analysis . . 59

7.4 Proposed Conceptual Stability Analysis 61

7.4.1 Stability Based Topics Selection 61

7.4.2 Conceptual Stability Computation 61

7.5 Published Baselines 63

7.5.1 Two NMF-based approaches 63

7.5.2 Four LDA-based approaches 64

7.6 Evaluation Metric 65

7.7 Datasets 65

7.8 Results and Comparison 66

7.8.1 Results on the First Group of Datasets 66

7.8.2 Results on the Second Group of Datasets 67

7.8.3 Remarks and Discussion 70

8 Discovering Hierarchy of Topics from Twitter Data 73

8.1 Previous Work on Twitter's Topic Modeling 74

8.2 A General Framework for Hierarchical Topic Discovery and Spatial Distribution 75

8.3 Datasets 76

8.4 Experiment Setup 76

8.5 Topics Discovery 78

8.5.1 Topics Discovery at the First Level 78

8.5.2 Topics Discovery at the Second Level 78

8.6 Topics Labeling 81

8.7 Geography Distribution of Topics 82

8.8 Discussion 82

9 Conclusion 93

9.1 Discussion 94

9.1.1 User Geolocation Prediction Scenario 94

9.1.2 Real-time Geolocation Prediction Scenario 94

9.1.3 Conceptual Stability Based Topic Modeling 95

9.1.4 General Framework for Discovering Hierarchy of Topics from Twitter Data 95

9.2 Future Research Direction 96

9.2.1 Scalable and robust methods for geolocation using diverse sources 96

9.2.2 Deep learning for text mining and analytics 96

9.2.3 Real-time topic modeling for data streams 96

9.2.4 Supervised or semi-supervised learning methods for topic labeling 96

9.2.5 A Complete application for topics discovery and visualization . . 97

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

A Appendixes for GMF and CSA 99
A.1 GMF 100
A.2 CSA 101
A.2.1 Datasets 101
A.2.2 Results 102
References 105

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.