

UNIVERSITY OF TECHNOLOGY SYDNEY

DOCTORAL THESIS

---

# Multi-Author Document Decomposition Based on Authorship

---

*Author:*

Khaled ALDEBEI

*Principal Supervisor:*

Prof. Xiangjian HE

*Co-Supervisor:*

Dr. Wenjing JIA

*Co-Supervisor:*

Dr. Gengfa FANG

*A thesis submitted in fulfilment of the requirements  
for the degree of Doctor of Philosophy*

*in the*

Global Big Data Technologies Centre  
UNIVERSITY OF TECHNOLOGY SYDNEY

January 2018

# Declaration of Authorship

I, Khaled ALDEBEI, declare that this thesis entitled, ‘Multi-Author Document Decomposition Based on Authorship’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Production Note:  
Signature removed prior to publication.

Date: 16/01/2018

UNIVERSITY OF TECHNOLOGY SYDNEY

# *Abstract*

Global Big Data Technologies Centre

UNIVERSITY OF TECHNOLOGY SYDNEY

Doctor of Philosophy

## **Multi-Author Document Decomposition Based on Authorship**

by Khaled ALDEBEI

Decomposing a document written by more than one author into sentences based on authorship is of great significance due to the increasing demand for plagiarism detection, forensic analysis, civil law (i.e., disputed copyright issues) and intelligence issues that involves disputed anonymous documents. Among the existing studies for document decomposition, some were limited by specific languages, according to topics or restricted to a document of two authors, and their accuracies have big rooms for improvement. In this thesis, we propose novel approaches for decomposition of a multi-author document written in any language disregarding to topics, based on a Naive-Bayesian model and Hidden Markov Model (HMM). The proposed approaches of the Naive-Bayesian model aim to exploit the difference in its posterior probability to improve the performance of decomposition. Two main procedures are proposed based on Naive-Bayesian model, and they are Segment Elicitation procedure and Probability Indication Procedure. The segment elicitation procedure is proposed to form a strong labeled training dataset. The probability indication procedure is developed to improve the purity of the sentence decomposition. The proposed approaches of the HMM strive to exploit the contextual correlation hidden among sentences when determining their authorships. In this thesis, it is for the first time the sequential patterns hidden among document elements is considered for such a problem. To build and learn the HMM, a new unsupervised learning method is proposed to estimate its initial parameters. The proposed frameworks do not require the availability of any information of authors or document's context other than

how many authors have contributed to writing the document. The effectiveness of the proposed algorithms is proved using benchmark datasets which are widely used for authorship analysis of documents. Furthermore, scientific papers are used to demonstrate the performance of the proposed approaches on authentic documents. Comparisons with recent state-the-art approaches are also presented to demonstrate the significance of our new ideas and the superior performance of the proposed approaches.

## *Acknowledgements*

Sincere feelings and strongest kind words emanating from my heart go to my supervisor **Professor Xiangjian He** for his marvelous support, advice, help and practical suggestions throughout my PhD journey. I am greatly indebted to him for his excellent direction, patient instruction, steadfast encouragement and timely comments to this research throughout the past four years. I owe my research achievements to his excellent supervision. Without his support and supervision, I could not have come this far.

I would like to thank my co-supervisor, **Dr. Wenjing Jia** for her unflinching encouragement, support and her useful comments. Her valuable suggestions and guidance have been a great help and kept me moving ahead at a critical time. I would also like to thank my co-supervisor **Dr. Gengfa Fang** for his advice and assistance.

I am forever appreciate and thankful to my family for all of their love and encouragement. Special thanks to my father, Mr. Waleed Aldebei, and my mother, Mrs. Sameera Abu Alsamen, for their extraordinarily generous help during my study. Without their support, I would never have had a chance to study overseas.

I would also like to express my pure-hearted thanks to my wife, Eng. Alanoud Alnsour, for her love, understanding and patience. You have been a pillar of strength to me. I also thank my darling baby girl, Salma Aldebei, for entering into our lives and making it more fulfilling.

Finally, I would like to extend my thanks to all my colleagues, friends and the staff of the school, especially those people listed below for providing support and friendship that I needed.

- Professor Massimo Piccardi, Associate Professor Qiang Wu, Dr. Min Xu, Dr. Priydarsi Nanda, Shaukat Abedi, Sari Awwad, Ahmed Mian Jan, Minqi Li, Fairouz Hussein, Omar Alshaweesh and Muhammad Usman.

# Contents

<b>Declaration of Authorship</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>Abbreviations</b>	<b>xiv</b>
<b>Publications</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background	1
1.2 Motivations and Objectives	4
1.3 Thesis Contributions	7
1.4 Thesis Structure	8
1.5 Summary	10
<b>2 Background and Related Work</b>	<b>11</b>
2.1 Authorship Analysis	11
2.1.1 Authorship Analysis Categories	14
2.1.1.1 Authorship Attribution	14
2.1.1.2 Authorship Verification	14
2.1.1.3 Plagiarism Detection	15
2.1.1.4 Authorship Profiling	16
2.1.1.5 Authorship-Based Text Decomposition	16
2.1.2 Stylometric Features	19
2.1.2.1 Syntactic Features	20
2.1.2.2 Lexical Features	21
2.1.2.3 Application Specific Features	23
2.1.3 Feature Representation	23

2.1.4	Approaches for Authorship Analysis . . . . .	24
2.2	Naive Bayes . . . . .	28
2.2.1	Bayes' Theorem . . . . .	28
2.2.2	Naive Bayesian Classifier . . . . .	29
2.2.2.1	Class Prior Probability . . . . .	31
2.2.2.2	Likelihood Probability . . . . .	32
2.2.3	Naive Bayesian in Document Analysis . . . . .	34
2.3	Sequential Data: Hidden Markov Model . . . . .	37
2.3.1	Markov Models . . . . .	38
2.3.2	Hidden Markov Model . . . . .	40
2.3.3	The Forward-Backward Algorithm . . . . .	43
2.3.4	The Viterbi Algorithm . . . . .	45
2.3.5	Hidden Markov Model in Document Analysis . . . . .	47
2.4	Clustering Methods: Gaussian Mixture Models . . . . .	48
2.4.1	The Gaussian Distribution . . . . .	49
2.4.2	Mixtures of Gaussians . . . . .	50
2.4.3	Expectation-Maximisation for GMMs . . . . .	51
2.4.4	Gaussian Mixture Models in Document Analysis . . . . .	52
2.5	Summary . . . . .	53
<b>3</b>	<b>Unsupervised Decomposition of a Multi-Author Document Based on Naive-Bayesian Model</b> . . . . .	<b>55</b>
3.1	Introduction . . . . .	56
3.2	Framework of the Proposed Approach . . . . .	57
3.3	Segmentation, Feature Representations and Clustering . . . . .	60
3.4	Segment Elicitation Procedure and Feature Re-vectorization . . . . .	63
3.5	Supervised Learning . . . . .	65
3.6	Probability Indication Procedure . . . . .	66
3.7	Experiments . . . . .	68
3.7.1	Datasets . . . . .	70
3.7.2	Experimental Results . . . . .	71
3.7.2.1	Results on Becker-Posner Blogs Dataset (Controlling for Topic) . . . . .	72
3.7.2.2	Results on <i>New York Times</i> Articles Dataset ( $N \geq 2$ ) . . . . .	72
3.7.2.3	Results on the Biblical Books Dataset . . . . .	73
3.7.2.4	Results on Authentic Document . . . . .	75
3.8	Summary . . . . .	76
<b>4</b>	<b>An Unsupervised Hierarchical Framework for Authorship-based Segmentation of a Multi-Author Document</b> . . . . .	<b>77</b>
4.1	Introduction . . . . .	78
4.2	Framework of the Proposed Approach . . . . .	81
4.3	First Level Learning . . . . .	83
4.3.1	Segmentation, Feature Extraction and Clustering . . . . .	83
4.3.2	Modified Segment Elicitation Procedure . . . . .	84
4.3.3	First-Stage Classification . . . . .	86
4.4	Second Level Learning . . . . .	87

4.4.1	Generating Training Dataset of Sentences . . . . .	87
4.4.2	Second-Stage Classification . . . . .	89
4.4.3	Final Refinement . . . . .	89
4.5	Experiments . . . . .	90
4.5.1	Datasets . . . . .	90
4.5.2	Experimental Results . . . . .	91
4.5.2.1	Results on the Becker-Posner Blogs Dataset ( <i>Controlling for Topic</i> ) . . . . .	91
4.5.2.2	Results on <i>New York Times</i> Articles Dataset ( $N \geq 2$ ) . . . . .	92
4.5.2.3	Results on Scientific Document . . . . .	96
4.6	Summary . . . . .	96
<b>5</b>	<b>Unsupervised Multi-Author Document Decomposition Based on Hidden Markov Model</b> . . . . .	<b>98</b>
5.1	Introduction . . . . .	99
5.2	Framework of the Proposed Approach . . . . .	102
5.3	Initializing Parameters of HMM . . . . .	103
5.4	Learning HMM . . . . .	105
5.5	Viterbi Decoding . . . . .	106
5.6	Experiments . . . . .	106
5.6.1	Datasets . . . . .	106
5.6.2	Experimental Results on Document Decomposition . . . . .	107
5.6.2.1	Results on the Biblical Books Dataset . . . . .	107
5.6.2.2	Results on Becker-Posner Blogs Dataset ( <i>Controlling for Topic</i> ) . . . . .	109
5.6.2.3	Results on <i>New York Times</i> Articles Dataset ( $N \geq 2$ ) . . . . .	110
5.6.2.4	Results on Scientific Document . . . . .	111
5.6.3	Experimental Results on Authorship Attribution . . . . .	112
5.7	Summary . . . . .	114
<b>6</b>	<b>SUDMAD: Sequential and Unsupervised Decomposition of a Multi-Author Document Based on a Hidden Markov Model</b> . . . . .	<b>116</b>
6.1	Introduction . . . . .	117
6.2	Framework of the Proposed SequentialUD Approach . . . . .	121
6.3	Estimating a Preliminary HMM from Unlabelled Input Data . . . . .	124
6.3.1	Hidden Markov Model . . . . .	124
6.3.2	Estimating Initial Parameters of HMM . . . . .	125
6.3.2.1	Estimating Transition Matrix $\mathbf{A}$ . . . . .	126
6.3.2.2	the Prior $\boldsymbol{\pi}$ . . . . .	128
6.3.2.3	Estimating the Emission Probabilities $\mathbf{B}$ . . . . .	128
6.3.3	Learning the Preliminary HMM . . . . .	129
6.3.4	Initial Sentence Decoding . . . . .	130
6.4	Learning the Boosted HMM . . . . .	131
6.4.1	Creating Consecutive-Sentence Dataset . . . . .	131
6.4.2	Re-Estimating and Learning the HMM parameters, and Final-Stage Sentence Decoding . . . . .	132
6.5	Refinement with ModPIP . . . . .	133
6.6	Experiments . . . . .	135



---

6.6.1	Datasets . . . . .	135
6.6.2	Experimental Results . . . . .	136
6.6.2.1	Results on the Biblical Books Dataset . . . . .	141
6.6.2.2	Results on Becker-Posner Blogs Dataset (Controlling for Topic) . . . . .	143
6.6.2.3	Results on <i>New York Times</i> Articles Dataset ( $N \geq 2$ ) . .	146
6.6.2.4	Results on Randomly Selected Scientific Articles . . . . .	152
6.6.2.5	Results on <i>Sanditon: An Unfinished Novel</i> . . . . .	153
6.6.2.6	Results on Scientific Document . . . . .	154
6.7	Summary . . . . .	156
<b>7</b>	<b>Conclusions</b>	<b>157</b>
	<b>Bibliography</b>	<b>160</b>

# List of Figures

1.1	An illustration of the decomposing process of a document written by $N$ authors. . . . .	3
2.1	An illustration of the supervised learning of text data. . . . .	27
2.2	An illustration of the unsupervised learning of text data. . . . .	27
2.3	The conditional independent assumption of features in vector $x = \{x^1, x^2, \dots, x^D\}$ given the class $y_j$ . . . . .	32
2.4	A representation of $N$ sequential data represented as independent, corresponding to a graph without links. . . . .	38
2.5	The first-order Markov chain. . . . .	39
2.6	The second-order Markov chain. . . . .	40
2.7	A graphical model of the HMM with $N$ hidden states, $Q = \{q_1, q_2, \dots, q_N\}$ , and $N$ observations, $X = \{x_1, x_2, \dots, x_N\}$ . . . . .	42
3.1	The framework of the proposed approach . . . . .	59
3.2	The illustration of criteria 2 and 3 of the probability indication procedure. $TS_{c_i}$ and $TS_{c_j}$ are trusted sentences for classes $c_i$ and $c_j$ , respectively. . . . .	67
3.3	The illustration of criterion 4 of the probability indication procedure. $TS_{c_i}$ is a trusted sentence for class $c_i$ . . . . .	67
3.4	The illustration of criterion 5 of the probability indication procedure. $TS_{c_i}$ and $TS_{c_j}$ are trusted sentences for classes $c_i$ and $c_j$ , respectively. . . . .	68
3.5	Purity results of the approaches proposed by Akiva and Koppel (2012), Akiva and Koppel (2013) and our proposed approach using documents created by three or four <i>New York Times</i> authors. . . . .	74
4.1	The proposed two-level, unsupervised learning framework. . . . .	82
4.2	Comparison of the purity results obtained using our approach and the approach in Giannella (2015) on the six single-topic documents. . . . .	93
4.3	Comparison of the purity results obtained using the <i>Proposed-1</i> approach and our approach on the six documents created by merging <i>New York Times</i> articles of two columnists. . . . .	94
5.1	Comparisons between using segments and using sentences in the unsupervised method for estimating the initial values of the HMM of our approach in terms of purity (represented as the cylinders) and number of iterations required for convergence (represented as the numbers above cylinders) using the 10 merged Bible documents. . . . .	109

5.2	Purity comparisons between our approach and the approaches presented in Akiva and Koppel (2013) and <i>Proposed-1</i> in Becker-Posner documents, and documents created by three or four <i>New York Times</i> columnists (TF = Thomas Friedman, PK = Paul Krugman, MD = Maureen Dowd, GC = Gail Collins).	110
5.3	Purity comparisons between our approach and the approach presented in (Giannella, 2015) in the six single-topic documents of Becker-Posner blogs.	111
6.1	The framework of the proposed SequentialUD and its refined version.	123
6.2	The recall rates of the clustering process obtained using words that have occurred at least once, twice, three times, four times and five times in four documents as features of BagOfWords1.	141
6.3	Purity results achieved on Becker-Posner blogs when our SequentialUD and its refined version are applied using different values of the limitation used to create the consecutive-sentence dataset.	144
6.4	Purity results of the approaches proposed by Giannella (2015), our SequentialUD and our refined SequentialUD using the six single-topic documents of Becker-Posner blogs.	145
6.5	Purity results obtained by using the merged documents of Becker-Posner blogs created by assigning different values of $V$ using our proposed approach SequentialUD and its refined version.	146
6.6	Purity results of the approaches proposed by Akiva and Koppel (2012), Akiva and Koppel (2013), <i>Proposed-1</i> , SequentialUD and refined SequentialUD using documents composed by merging articles of three and four <i>New York Times</i> columnists.	148
6.7	Purity results of the refined SequentialUD approach with respect to the maximum number of sentences located in a group, as indicated in Criteria 4 and 5 of the ModPIP, using the document written by four <i>New York Times</i> columnists when different values of threshold $R$ (the horizontal axis) are used in the ModPIP. In the graph, the numbers above the line markers indicate the maximum number of sentences located in a group.	149
6.9	Purity results of SequentialUD approach and its refined version on merged documents created by merging 1, 5, 10, 15 and 20 randomly selected articles of two, three and four authors. The error bars depict 0.95 confidence interval for the refined SequentialUD approach. In many cases the confidence intervals are quite small and are not easily seen in the figure.	151
6.8	Comparison of the purity results obtained using the approach in Giannella (2015), SequentialUD approach and refined SequentialUD approach on short documents with short consecutive sentences composed by merging articles of four <i>New York Times</i> columnists using the same procedure of Giannella (2015), when the mean author run length (i.e., meanARL) is varied. The error bars depict 0.95 confidence interval for the three approaches.	152

# List of Tables

3.1	The clustering results of segments in the Ezekiel-Job document . . . . .	61
3.2	Purity results of applying our approach in the Ezekiel-Job document using different values of segment length ( $v$ ) and different values of vital segments percentage ( $s$ ) . . . . .	62
3.3	The purity results of sentences in the Ezekiel-Job document . . . . .	66
3.4	The classified sentences and correctly classified sentences of the Ezekiel-Job document by applying the five criteria of the probability indication procedure . . . . .	69
3.5	The purity results obtained by using different values of $q$ in Criterion 1 of the probability indication procedure on the Ezekiel-Job document . . . . .	69
3.6	Statistics of the <i>New York Times</i> articles. . . . .	70
3.7	Statistics regarding the five Bible books. . . . .	71
3.8	Purity comparison on a document of Becker-Posner Blogs. Approaches compared: 1- Akiva and Koppel (2012), 2- Akiva and Koppel (2013) and 3- Our approach. . . . .	72
3.9	The purity results of documents created by merging any pair of the four <i>New York Times</i> columnists using our proposed approach. . . . .	73
3.10	Purity comparison on documents composed by merging two biblical books of <i>different literatures</i> . Approaches in comparison: 1- Koppel et al. (2011a), 2- Akiva and Koppel (2013), 3- Akiva and Koppel (2013)-SynonymSet, 4- Our Proposed Approach. . . . .	74
3.11	Purity comparison on documents composed by merging two biblical books of the <i>same genre</i> . Approaches in comparison: 1- Koppel et al. (2011a), 2- Akiva and Koppel (2012), 3- Akiva and Koppel (2013), 4- Akiva and Koppel (2013)-SynonymSet, 5- Our Proposed Approach. . . . .	75
4.1	Statistics of the six single-topic documents created from the Becker-Posner Blogs. . . . .	91
4.2	Purity results of the document of all Becker-Posner blogs using the approaches of [1] Akiva and Koppel (2012), [2] Akiva and Koppel (2013), [3] <i>Proposed-1</i> , [4] First level learning and our approach. . . . .	92
4.3	Purity results of the documents merged from the articles written by three or four of the <i>New York Times</i> columnists, respectively, using the approaches of [1] Akiva and Koppel (2012), [2] Akiva and Koppel (2013), [3] <i>Proposed-1</i> and our approach. . . . .	94
4.4	Purity results of documents created by merging two bibles of different literatures. Approaches in comparison: [1] Koppel et al. (2011a), [2] Akiva and Koppel (2013), [3] Akiva and Koppel (2013)-SynonymSet, [4] <i>Proposed-1</i> and our approach. . . . .	95

4.5	Purity results of documents created by merging two bibles of different literatures. Approaches in comparison are noted as: [1] Koppel et al. (2011a), [2] Akiva and Koppel (2012), [3] Akiva and Koppel (2013), [4] Akiva and Koppel (2013)-SynonymSet, [5] <i>Proposed-1</i> and our approach. . . . .	96
4.6	The purity results and predicted contributions of two authors of a scientific paper obtained using the proposed approach. . . . .	97
5.1	Purity results of merged documents of <i>different literature</i> bible books using the approaches of 1- Koppel et al. (2011a), 2- Akiva and Koppel (2013)-500CommonWords, 3- Akiva and Koppel (2013)-SynonymSet, 4- <i>Proposed-1</i> and 5- our approach. . . . .	108
5.2	Purity results of merged documents of the <i>same literature</i> bible books using the approaches of 1- Koppel et al. (2011a), 2- Akiva and Koppel (2012), 3- Akiva and Koppel (2013)-500CommonWords, 4- Akiva and Koppel (2013)-SynonymSet, 5- <i>Proposed-1</i> and 6- our approach. . . . .	108
5.3	The purity results and predicted contributions of the two authors of the scientific paper using the proposed approach. . . . .	112
5.4	The number of sentences that are classified with Madison sentences and Hamilton sentences of each of the 12 anonymous articles of <i>The Federalist Papers</i> using the proposed approach. . . . .	114
6.1	Purity results of applying our SequentialUD approach on the selected Eze-Prov document (a Long Document) with different $v$ and meanARL. Note that better purity results (highlighted in bold font) are achieved when $v$ is less than meanARL and 60. . . . .	139
6.2	Purity results of applying our SequentialUD approach on the scientific document (a Short Document) with different $v$ and meanARL. Note that better purity results (highlighted in bold font) are achieved when $v$ is less than meanARL and 40. . . . .	140
6.3	Purity comparison on documents composed by merging two biblical books of <i>different genres</i> . Approaches in comparison: 1- Koppel et al. (2011a), 2- Akiva and Koppel (2013), 3- Akiva and Koppel (2013)-SynonymSet, 4- <i>Proposed-1</i> , 5- Our SequentialUD and 6- Our refined SequentialUD. . . . .	142
6.4	Purity comparison on documents composed by merging two biblical books of the <i>same genre</i> . Approaches in comparison: 1- Koppel et al. (2011a), 2- Akiva and Koppel (2012), 3- Akiva and Koppel (2013), 4- Akiva and Koppel (2013)-SynonymSet, 5- <i>Proposed-1</i> , 6- Our SequentialUD and 7- Our refined SequentialUD. . . . .	142
6.5	Purity comparison on a document of Becker-Posner Blogs. Approaches compared: 1- Akiva and Koppel (2012), 2- Akiva and Koppel (2013), 3- <i>Proposed-1</i> , 4- First-Stage HMM, 5-Our SequentialUD and 6- Our Refined SequentialUD. . . . .	143
6.6	The purity results of documents created by merging any pair of the four <i>New York Times</i> columnists using the <i>Proposed-1</i> approach, our SequentialUD and our refined SequentialUD. . . . .	147
6.7	The purity results and predicted contributions of the two authors of the scientific paper using the proposed approach SequentialUD and its refined version. . . . .	155

---

6.8	The maximum number of sentences that are located in groups regarding criteria 4 and 5 of the ModPIP using corpus used in this article when the value of threshold R is equal to 15. . . . .	155
-----	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----

# Abbreviations

<b>AA</b>	<b>A</b> uthorship <b>A</b> ttribution
<b>DT</b>	<b>D</b> ecision <b>T</b> ree
<b>EM</b>	<b>E</b> xpectation <b>M</b> aximisation
<b>GMM</b>	<b>G</b> aussian <b>M</b> ixture <b>M</b> odel
<b>HMM</b>	<b>H</b> idden <b>M</b> arkov <b>M</b> odel
<b>IID</b>	<b>I</b> ndependent and <b>I</b> dentically <b>D</b> istributed
<b>IR</b>	<b>I</b> nformation <b>R</b> etrieval
<b>K-NN</b>	<b>K</b> -Nearest <b>N</b> eighbors
<b>MAP</b>	<b>M</b> aximum <b>A</b> <b>P</b> osterior
<b>MLE</b>	<b>M</b> aximum <b>L</b> ikelihood <b>E</b> stimation
<b>NN</b>	<b>N</b> eural <b>N</b> etworks
<b>NLP</b>	<b>N</b> atural <b>L</b> anguage <b>P</b> rocessing
<b>NUANCE</b>	<b>N</b> -gram <b>U</b> nsupervised <b>A</b> utomated <b>N</b> atural <b>C</b> luster <b>E</b> nsemble
<b>PIP</b>	<b>P</b> robability <b>I</b> ndication <b>P</b> rocedure
<b>POS</b>	<b>P</b> art- <b>O</b> f- <b>S</b> peech
<b>SUDMAD</b>	<b>S</b> equential and <b>U</b> nsupervised <b>D</b> ecomposition of a <b>M</b> ulti- <b>A</b> uthor <b>D</b> ocument
<b>SVM</b>	<b>S</b> upport <b>V</b> ector <b>M</b> achine
<b>TF-IDF</b>	<b>T</b> erm <b>F</b> requency- <b>I</b> nverse <b>D</b> ocument <b>F</b> requency
<b>WAN</b>	<b>W</b> ord <b>A</b> gency <b>N</b> etwork

# Publications

These are the publications resulted from this thesis.

- Khaled Aldebei, Xiangjian He, and Jie Yang. Unsupervised Decomposition of a Multi-Author Document Based on Naive-Bayesian Model. *Association for Computational Linguistics, Volume 2: Short Papers*, page 501, 2015. (**CORE A\***, **CCF A & ERA A**)
- Khaled Aldebei, Xiangjian He, Wenjing Jia, and Jie Yang. Unsupervised Multi-Author Document Decomposition Based on Hidden Markov Model. In *ACL (1)*, 2016a. (**CORE A\***, **CCF A & ERA A**)
- Khaled Aldebei, Helia Farhood, Wenjing Jia, Priyadarsi Nanda, and Xiangjian He. Sequential and Unsupervised Document Authorial Clustering Based on Hidden Markov Model. In *Trustcom/BigDataSE/ICCESS, 2017 IEEE*, pages 379–385. IEEE, 2017. (**CORE A & ERA A**)
- Khaled Aldebei, Xiangjian He, Wenjing Jia, and Weichang Yeh. SUDMAD: Sequential and Unsupervised Decomposition of a Multi-Author Document Based on a Hidden Markov Model. *Journal of the Association for Information Science and Technology*, 69(2):201–214, 2018. ISSN 2330-1643. doi: 10.1002/asi.23956. URL <http://dx.doi.org/10.1002/asi.23956>. (**ERA A\* & CCF B**)



*Dedicated to My Family*