

## TABLE OF CONTENTS

<b>RELATED BOOKS BY THE AUTHOR .....</b>	<b>XV</b>
<b>CHAPTER 1 –WHAT IS NLP ALL ABOUT? .....</b>	<b>1</b>
1.1 WHAT IS NATURAL LANGUAGE PROCESSING? .....	1
1.2 OBJECTIONS .....	3
1.3 WHO ELSE CARES? .....	4
1.4 IMPORTANCE NATURAL LANGUAGE PROCESSING .....	4
1.5 NATURAL LANGUAGE PROCESSING APPLICATIONS.....	5
1.5.1 <i>Compliance</i> .....	5
1.5.2 <i>Customer Satisfaction</i> .....	5
1.5.3 <i>Customer Complaints</i> .....	6
1.5.4 <i>Document Analysis</i> .....	7
1.5.5 <i>Tweet Sentiment</i> .....	7
1.6 SUMMARY .....	11
1.7 CODE INPUTS AND OUTPUTS .....	12
1.7.1 <i>Python with Jupyter Notebooks</i> .....	12
1.7.2 <i>R with RStudio</i> .....	14
1.7.3 <i>Basic Online Tutorials</i> .....	15
<b>CHAPTER 2 –INTRODUCTION TO TEXT MINING .....</b>	<b>17</b>
2.1 INTRODUCTION .....	17
2.2 LOAD THE R PACKAGES .....	17
2.3 CORPUS PREPROCESSING .....	18
2.4 LOAD THE DATA .....	21
2.4.1 <i>Loading Data in Python</i> .....	21
2.4.2 <i>Lodaing Data in R</i> .....	23
2.4.3 <i>Loading Our Data</i> .....	25
2.5 CONVERT TO LOWERCASE .....	27
2.5.1 <i>Convert to Lowercase in Python</i> .....	27
2.5.2 <i>Convert to Lowercase in R</i> .....	28
2.6 REMOVE NUMBERS, PUNCTUATION, & WHITESPACES.....	29
2.6.1 <i>Remove Numbers in R</i> .....	29
2.6.2 <i>Remove Punctuation in R</i> .....	29
2.6.3 <i>Remove Whitespaces in R</i> .....	29
2.6.4 <i>Remove Numbers in Python</i> .....	30
2.7 TOKENIZE THE TEXT .....	31
2.7.1 <i>Tokenization in Python</i> .....	31
2.7.2 <i>Tokeniztion in R</i> .....	31
2.8 WORD AND NUMBER REMOVAL.....	32
2.8.1 <i>Remove Stopwords in R</i> .....	32
2.8.2 <i>Removing Stopwords in Python</i> .....	32
2.8.3 <i>Removing Stopwords from Our Text</i> .....	33

2.9	STEMMING THE TEXT.....	33
2.9.1	<i>Stemming in R</i> .....	34
2.9.2	<i>Stemming in Python</i> .....	34
2.9.3	<i>Stemming our text</i> .....	34
2.10	TEXT LEMMATIZATION .....	34
2.10.1	<i>Lemmatization in Python</i> .....	35
2.10.2	<i>Lemmatization in R</i> .....	35
2.11	DOCUMENT MATRIX CREATION.....	37
2.11.1	<i>Create Document Matrices</i> .....	37
2.11.2	<i>Develop Frequency Distributions</i> .....	38
2.11.3	<i>Visualizing the Results</i> .....	39
2.11.4	<i>Find correlations</i> .....	40
2.11.5	<i>Using Wordclouds to Visualize Results</i> .....	41
2.12	TIDY TEXT ANALYTICS I .....	42
2.12.1	<i>Tidy Format</i> .....	42
2.12.2	<i>Tidy Text Format</i> .....	42
2.12.3	<i>Contrasting tidy text with other data structures</i> .....	42
2.12.4	<i>The unnest_tokens function</i> .....	43
2.12.5	<i>Tibbles</i> .....	44
2.12.6	<i>Tokenization</i> .....	45
2.13	EXAMPLE - INDIAN PHILOSOPHY .....	46
2.13.1	<i>Filter for negative words</i> .....	51
2.13.2	<i>Build a cloud chart</i> .....	52
2.14	TIDY TEXT ANALYTICS II .....	54
2.14.1	<i>Tidy Text</i> .....	54
2.14.2	<i>Quantum Words</i> .....	55
2.14.3	<i>Quantum Tokens</i> .....	55
2.14.4	<i>Quantum Cloud</i> .....	56
2.14.5	<i>Lexicon Exploration</i> .....	57
2.14.6	<i>Loughran Lexicon</i> .....	58
2.14.7	<i>AFINN Lexicon</i> .....	58
2.14.8	<i>Bing Lexicon</i> .....	59
2.14.9	<i>Getting Sentiments with Loughran</i> .....	60
2.14.10	<i>Analyzing word and document frequency: tf-idf</i> .....	60
2.14.11	<i>Term frequency</i> .....	62
2.14.12	<i>Get Term Frequencies</i> .....	62
2.14.13	<i>Zipf's law</i> .....	64
2.14.14	<i>The bind_tf_idf function</i> .....	67
2.14.15	<i>Summary</i> .....	71
2.15	EXERCISES .....	71
<b>CHAPTER 3 –VECTORIZING TEXT DATA.....</b>		<b>75</b>
3.1	INTRODUCTION .....	75
3.2	DEFINITIONS .....	75

3.3	TYPES OF WORD EMBEDDINGS .....	76
3.3.1	<i>Frequency-based Embedding</i> .....	76
3.3.2	<i>Prediction-based Embedding</i> .....	81
3.4	PLAIN TEXT TECHNICAL DETAILS .....	81
3.5	WORD VECTORIZING WITH PYTHON .....	82
3.6	SUMMARY .....	87
3.7	EXERCISE .....	88
<b>CHAPTER 4 –VECTORIZING DATA APPLICATION .....</b>		<b>89</b>
4.1	INTRODUCTION¶ .....	89
4.2	METHODOLOGY¶ .....	90
4.3	LOAD AND READING DATA¶ .....	91
4.3.1	<i>Loading (reading) the Dataset using Pandas</i> .....	92
4.4	DATA EXPLORATION.....	92
4.4.1	<i>Prepare a Frequency Distribution</i> .....	93
4.4.2	<i>Plotting the Data</i> .....	94
4.4.3	<i>Histogram with Legend using Matplotlib Only</i> .....	95
4.4.4	<i>Store Product Group Data</i> .....	96
4.4.5	<i>Define Variables</i> .....	97
4.4.6	<i>Label Encoding of Classes:</i> .....	97
4.4.7	<i>Word Cloud Visualization:</i> .....	99
4.5	NATURAL LANGUAGE PROCESSING (NLP) .....	103
4.5.1	<i>Stop Words</i> .....	103
4.5.2	<i>Customized Stop Words</i> .....	104
4.5.3	<i>Lemmatization</i> .....	105
4.5.4	<i>Create Train and Test Sets</i> .....	105
4.5.5	<i>Feature Extraction</i> .....	106
4.5.6	<i>Vectorization</i> .....	106
4.5.7	<i>CountVectorizer</i> .....	106
4.5.8	<i>Term Frequency–Inverse Document Frequency</i> .....	109
4.6	DISCRETE CLASSIFIERS .....	109
4.6.1	<i>Naïve Bayes Classification</i> .....	110
4.6.2	<i>Pipeline Definition</i> .....	111
4.6.3	<i>Product-to-Category and Category-to-Product Definition</i> .....	112
4.6.4	<i>TfidfVectorizer</i> .....	113
4.6.5	<i>Extracting N-Grams</i> .....	115
4.7	BAG-OF-WORDS (BOW) PREPARATION .....	118
4.7.1	<i>About Bag of Words</i> .....	118
4.7.2	<i>Feature Engineering using Bag-of-Words</i> .....	118
4.7.3	<i>Vectorizing the Complaints</i> .....	118
4.8	BOW WITH A MULTINOMIAL BAYES CLASSIFIER.....	119
4.8.1	<i>Classifier Training Score</i> .....	119
4.8.2	<i>Classifier Prediction</i> .....	120
4.8.3	<i>Classifier Testing Score</i> .....	120

4.8.4	<i>Confusion Matrices</i> .....	120
4.8.5	<i>Normalized Confusion Matrix</i> .....	122
4.8.6	<i>Performance Metrics</i> .....	123
4.8.7	<i>Performance</i> .....	123
4.9	CLASSIFIER COMPARISON .....	123
4.9.1	<i>Refresh Train and Test BOWs</i> .....	124
4.9.2	<i>Define Classifiers</i> .....	125
4.9.3	<i>Execute the Classifiers</i> .....	125
4.10	MODEL ANALYSIS .....	127
4.10.1	<i>Analysis of Model Performance</i> .....	127
4.10.2	<i>BOW Analysis with a Multinomial Logistic Regression</i> .....	128
4.10.3	<i>Confusion Matrix</i> .....	128
4.10.4	<i>Normalized Confusion Matrix</i> .....	130
4.10.5	<i>Classification Report (Precision, Recall and F1-Score)</i> .....	133
4.10.6	<i>Conclusion</i> .....	133
4.11	METRICS DEFINED .....	133
4.11.1	<i>Accuracy</i> .....	133
4.11.2	<i>Precision</i> .....	134
4.11.3	<i>Recall</i> .....	134
4.11.4	<i>F1 Measure</i> .....	135
4.12	RIDGE REGRESSION CLASSIFIER.....	135
4.13	EXERCISES .....	138
<b>CHAPTER 5 –TWEET SENTIMENT ANALYSIS .....</b>		<b>141</b>
5.1	INTRODUCTION .....	141
5.2	SENTIMENT LEXICONS.....	141
5.3	INSTALL REQUIRED LIBRARIES .....	142
5.4	LEXICON EXAMPLE .....	143
5.5	EXAMPLE 1: THE INNER JOIN .....	144
5.5.1	<i>Using the Inner_Join to Analyze Sentiment</i> .....	146
5.5.2	<i>Positive &amp; Negative Words over time</i> .....	148
5.5.3	<i>Most common positive and negative words</i> .....	152
5.5.4	<i>Wordclouds</i> .....	154
5.5.5	<i>Word Frequencies</i> .....	156
5.5.6	<i>Comparing word usage</i> .....	159
5.5.7	<i>Changes in word use</i> .....	161
5.5.8	<i>Evaluating Changes using Slope</i> .....	164
5.6	EXAMPLE 2: ISISFANBOY .....	165
5.6.1	<i>Loading the Data</i> .....	166
5.6.2	<i>Fixing the dates</i> .....	167
5.6.3	<i>Tokenization</i> .....	167
5.6.4	<i>Removing common words</i> .....	168
5.6.5	<i>Word frequencies and word clouds</i> .....	169
5.6.6	<i>Plotting the most occurring positive and negative words</i> .....	170

5.6.7	<i>New Bigram Counts</i> .....	172
5.6.8	<i>Word clouds on positive and negative sentiment</i> .....	173
5.6.9	<i>Top 10 words contributing to different sentiments</i> .....	174
5.6.10	<i>Sentiment proportion analysis</i> .....	177
5.6.11	<i>Get AFINN Setiments</i> .....	178
5.6.12	<i>N-grams</i> .....	178
5.6.13	<i>Get Sentiment Scores</i> .....	180
5.6.14	<i>Network of bigrams</i> .....	181
5.7	EXERCISES .....	183
<b>CHAPTER 6 –LATENT DIRICHLET ALLOCATION (LDA)</b> .....		<b>185</b>
6.1	BLOG TOPIC ANALYSIS USING LDA .....	185
6.1.1	<i>Set working Directory</i> .....	185
6.1.2	<i>Read files into a character vector</i> .....	186
6.1.3	<i>Create corpus from vector</i> .....	186
6.1.4	<i>Start Preprocessing</i> .....	187
6.1.5	<i>Stem document</i> .....	188
6.1.6	<i>Create document-term matrix</i> .....	189
6.1.7	<i>Load Topic Models Library</i> .....	190
6.1.8	<i>Set Parameters for Gibbs Sampling</i> .....	190
6.1.9	<i>Run LDA using Gibbs sampling</i> .....	191
6.1.10	<i>Write out Results</i> .....	191
6.2	2012 USA PRESIDENTIAL DEBATE USING LDA .....	192
6.2.1	<i>Loading and Installing Libraries</i> .....	192
6.2.2	<i>Get External Scripts</i> .....	193
6.2.3	<i>Load the Data</i> .....	193
6.2.4	<i>Determine Optimal Number of Topics</i> .....	194
6.2.5	<i>Run the Model</i> .....	195
6.2.6	<i>Plot the Topics Per Person &amp; Time</i> .....	195
6.2.7	<i>Plot the Topics Matrix as a Heatmap</i> .....	197
6.2.8	<i>Network of the Word Distributions Over Topics</i> .....	198
6.2.9	<i>Topic Distributions over Candidates</i> .....	199
6.2.10	<i>LDavis of Model</i> .....	201
6.2.11	<i>Build a New Model</i> .....	203
6.2.12	<i>Plot the Topics Per Person &amp; Location for New Data</i> .....	205
6.3	OPTIMAL_K.....	207
6.4	EXERCISES .....	208
<b>R GLOSSARY</b> .....		<b>211</b>
<b>PYTHON GLOSSARY</b> .....		<b>219</b>
<b>REFERENCES</b> .....		<b>223</b>
<b>INDEX</b> .....		<b>227</b>