Practical Natural Language Processing
A Comprehensive Guide to Building Real-World NLP Systems

Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®
# Table of Contents

Foreword ................................................................. xv
Preface ................................................................. xvii

## Part I. Foundations

1. NLP: A Primer ......................................................... 3
   - NLP in the Real World ........................................ 5
   - NLP Tasks ..................................................... 6
   - What Is Language? ............................................. 8
     - Building Blocks of Language ......................... 9
     - Why Is NLP Challenging? .............................. 12
   - Machine Learning, Deep Learning, and NLP: An Overview .................................................... 14
   - Approaches to NLP ........................................... 16
     - Heuristics-Based NLP .................................... 16
     - Machine Learning for NLP ............................ 19
     - Deep Learning for NLP .................................. 22
     - Why Deep Learning Is Not Yet the Silver Bullet for NLP .................................................. 28
   - An NLP Walkthrough: Conversational Agents ......................................................... 31
   - Wrapping Up .................................................. 33

2. NLP Pipeline .......................................................... 37
   - Data Acquisition .............................................. 39
   - Text Extraction and Cleanup ............................ 42
     - HTML Parsing and Cleanup ............................ 44
     - Unicode Normalization .................................. 45
     - Spelling Correction ....................................... 46

...
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>System-Specific Error Correction</td>
<td>47</td>
</tr>
<tr>
<td>Pre-Processing</td>
<td>49</td>
</tr>
<tr>
<td>Preliminaries</td>
<td>50</td>
</tr>
<tr>
<td>Frequent Steps</td>
<td>52</td>
</tr>
<tr>
<td>Other Pre-Processing Steps</td>
<td>55</td>
</tr>
<tr>
<td>Advanced Processing</td>
<td>57</td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>60</td>
</tr>
<tr>
<td>Classical NLP/ML Pipeline</td>
<td>62</td>
</tr>
<tr>
<td>DL Pipeline</td>
<td>62</td>
</tr>
<tr>
<td>Modeling</td>
<td>62</td>
</tr>
<tr>
<td>Start with Simple Heuristics</td>
<td>63</td>
</tr>
<tr>
<td>Building Your Model</td>
<td>64</td>
</tr>
<tr>
<td>Building THE Model</td>
<td>65</td>
</tr>
<tr>
<td>Evaluation</td>
<td>68</td>
</tr>
<tr>
<td>Intrinsic Evaluation</td>
<td>68</td>
</tr>
<tr>
<td>Extrinsic Evaluation</td>
<td>71</td>
</tr>
<tr>
<td>Post-Modeling Phases</td>
<td>72</td>
</tr>
<tr>
<td>Deployment</td>
<td>72</td>
</tr>
<tr>
<td>Monitoring</td>
<td>72</td>
</tr>
<tr>
<td>Model Updating</td>
<td>73</td>
</tr>
<tr>
<td>Working with Other Languages</td>
<td>73</td>
</tr>
<tr>
<td>Case Study</td>
<td>74</td>
</tr>
<tr>
<td>Wrapping Up</td>
<td>76</td>
</tr>
</tbody>
</table>

3. **Text Representation** ............................................................................. 81

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Space Models</td>
<td>84</td>
</tr>
<tr>
<td>Basic Vectorization Approaches</td>
<td>85</td>
</tr>
<tr>
<td>One-Hot Encoding</td>
<td>85</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>87</td>
</tr>
<tr>
<td>Bag of N-Grams</td>
<td>89</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>90</td>
</tr>
<tr>
<td>Distributed Representations</td>
<td>92</td>
</tr>
<tr>
<td>Word Embeddings</td>
<td>94</td>
</tr>
<tr>
<td>Going Beyond Words</td>
<td>103</td>
</tr>
<tr>
<td>Distributed Representations Beyond Words and Characters</td>
<td>105</td>
</tr>
<tr>
<td>Universal Text Representations</td>
<td>107</td>
</tr>
<tr>
<td>Visualizing Embeddings</td>
<td>108</td>
</tr>
<tr>
<td>Handcrafted Feature Representations</td>
<td>112</td>
</tr>
<tr>
<td>Wrapping Up</td>
<td>113</td>
</tr>
</tbody>
</table>
## Part II. Essentials

### 4. Text Classification

<table>
<thead>
<tr>
<th>Applications</th>
<th>121</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Pipeline for Building Text Classification Systems</td>
<td>123</td>
</tr>
<tr>
<td>A Simple Classifier Without the Text Classification Pipeline</td>
<td>125</td>
</tr>
<tr>
<td>Using Existing Text Classification APIs</td>
<td>126</td>
</tr>
<tr>
<td>One Pipeline, Many Classifiers</td>
<td>126</td>
</tr>
<tr>
<td>Naive Bayes Classifier</td>
<td>127</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>131</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>132</td>
</tr>
<tr>
<td>Using Neural Embeddings in Text Classification</td>
<td>134</td>
</tr>
<tr>
<td>Word Embeddings</td>
<td>134</td>
</tr>
<tr>
<td>Subword Embeddings and fastText</td>
<td>136</td>
</tr>
<tr>
<td>Document Embeddings</td>
<td>138</td>
</tr>
<tr>
<td>Deep Learning for Text Classification</td>
<td>140</td>
</tr>
<tr>
<td>CNNs for Text Classification</td>
<td>143</td>
</tr>
<tr>
<td>LSTMs for Text Classification</td>
<td>144</td>
</tr>
<tr>
<td>Text Classification with Large, Pre-Trained Language Models</td>
<td>145</td>
</tr>
<tr>
<td>Interpreting Text Classification Models</td>
<td>147</td>
</tr>
<tr>
<td>Explaining Classifier Predictions with Lime</td>
<td>148</td>
</tr>
<tr>
<td>Learning with No or Less Data and Adapting to New Domains</td>
<td>149</td>
</tr>
<tr>
<td>No Training Data</td>
<td>149</td>
</tr>
<tr>
<td>Less Training Data: Active Learning and Domain Adaptation</td>
<td>150</td>
</tr>
<tr>
<td>Case Study: Corporate Ticketing</td>
<td>152</td>
</tr>
<tr>
<td>Practical Advice</td>
<td>155</td>
</tr>
<tr>
<td>Wrapping Up</td>
<td>157</td>
</tr>
</tbody>
</table>

### 5. Information Extraction

<table>
<thead>
<tr>
<th>Applications</th>
<th>162</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE Tasks</td>
<td>164</td>
</tr>
<tr>
<td>The General Pipeline for IE</td>
<td>165</td>
</tr>
<tr>
<td>Keyphrase Extraction</td>
<td>166</td>
</tr>
<tr>
<td>Implementing KPE</td>
<td>167</td>
</tr>
<tr>
<td>Practical Advice</td>
<td>168</td>
</tr>
<tr>
<td>Named Entity Recognition</td>
<td>169</td>
</tr>
<tr>
<td>Building an NER System</td>
<td>171</td>
</tr>
<tr>
<td>NER Using an Existing Library</td>
<td>175</td>
</tr>
<tr>
<td>NER Using Active Learning</td>
<td>176</td>
</tr>
<tr>
<td>Practical Advice</td>
<td>177</td>
</tr>
<tr>
<td>Named Entity Disambiguation and Linking</td>
<td>178</td>
</tr>
<tr>
<td>NEL Using Azure API</td>
<td>179</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>8.</td>
<td>Social Media.</td>
</tr>
<tr>
<td></td>
<td>Applications</td>
</tr>
<tr>
<td></td>
<td>Unique Challenges</td>
</tr>
<tr>
<td></td>
<td>NLP for Social Data</td>
</tr>
<tr>
<td></td>
<td>Word Cloud</td>
</tr>
<tr>
<td></td>
<td>Tokenizer for SMTD</td>
</tr>
<tr>
<td></td>
<td>Trending Topics</td>
</tr>
<tr>
<td></td>
<td>Understanding Twitter Sentiment</td>
</tr>
<tr>
<td></td>
<td>Pre-Processing SMTD</td>
</tr>
<tr>
<td></td>
<td>Text Representation for SMTD</td>
</tr>
<tr>
<td></td>
<td>Customer Support on Social Channels</td>
</tr>
<tr>
<td></td>
<td>Memes and Fake News</td>
</tr>
<tr>
<td></td>
<td>Identifying Memes</td>
</tr>
<tr>
<td></td>
<td>Fake News</td>
</tr>
<tr>
<td></td>
<td>Wrapping Up</td>
</tr>
<tr>
<td>9.</td>
<td>E-Commerce and Retail.</td>
</tr>
<tr>
<td></td>
<td>E-Commerce Catalog</td>
</tr>
<tr>
<td></td>
<td>Review Analysis</td>
</tr>
<tr>
<td></td>
<td>Product Search</td>
</tr>
<tr>
<td></td>
<td>Product Recommendations</td>
</tr>
<tr>
<td></td>
<td>Search in E-Commerce</td>
</tr>
<tr>
<td></td>
<td>Building an E-Commerce Catalog</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>The Data Science Process</td>
<td>390</td>
</tr>
<tr>
<td>The KDD Process</td>
<td>390</td>
</tr>
<tr>
<td>Microsoft Team Data Science Process</td>
<td>392</td>
</tr>
<tr>
<td>Making AI Succeed at Your Organization</td>
<td>394</td>
</tr>
<tr>
<td>Team</td>
<td>394</td>
</tr>
<tr>
<td>Right Problem and Right Expectations</td>
<td>395</td>
</tr>
<tr>
<td>Data and Timing</td>
<td>396</td>
</tr>
<tr>
<td>A Good Process</td>
<td>397</td>
</tr>
<tr>
<td>Other Aspects</td>
<td>398</td>
</tr>
<tr>
<td>Peeking over the Horizon</td>
<td>400</td>
</tr>
<tr>
<td>Final Words</td>
<td>403</td>
</tr>
</tbody>
</table>

Index.............................................................................................................. 409